



ARTIFICIAL NEURAL NETWORK PREDICTION OF CHEMICAL-DISEASE
RELATIONSHIPS USING READILY AVAILABLE CHEMICAL PROPERTIES

THESIS

Edward J. Brouch, Captain, USAF

AFIT-ENV-14-M-12

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

DISTRIBUTION STATEMENT A.
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government.

AFIT-ENV-14-M-12

ARTIFICIAL NEURAL NETWORK PREDICTION OF CHEMICAL-DISEASE
RELATIONSHIPS USING READILY AVAILABLE CHEMICAL PROPERTIES

THESIS

Presented to the Faculty

Department of Systems Engineering and Management

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the
Degree of Master of Science in Engineering Management

Edward J. Brouch, BS

Captain, USAF

March 2014

DISTRIBUTION STATEMENT A.
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

ARTIFICIAL NEURAL NETWORK PREDICTION OF CHEMICAL-DISEASE
RELATIONSHIPS USING READILY AVAILABLE CHEMICAL PROPERTIES

Edward J. Brouch, BS
Captain, USAF

Approved:

//signed//
Willie Harper, PhD (Chairman)

20 February 2014
Date

//signed//
Tay Johannes, Lt Col, USAF (Member)

20 February 2014
Date

//signed//
Dirk Yamamoto, Lt Col, USAF (Member)

20 February 2014
Date

Abstract

The natural environment is burdened with a broad range of toxic chemicals, and there is a need to develop a tool that can accelerate the pace at which we learn how chemicals impact disease. This work developed an artificial neural network (ANN) based model that constructed chemical-disease relationships for chemicals found in the Comparative Toxicogenomics Database. A new chemical classification system, based on the molecular weight, hydrogen donors, and hydrogen acceptors, was created to identify chemicals with a unique number that is directly related to these structural properties of the chemical. Diseases were grouped into 27 categories and the chemical-disease associations were made between the chemical and its associated disease category. The ANN model was successfully trained and tested to associated 75 chemical with the 27 disease categories. Simulations with training-validation-testing ratios of 70-15-15 percent produced coefficients of determination equal to 0.99, and the Levenberg-Marquardt backpropagation function provided the best network performance. To help validate the model, the ANN was also used to evaluate chemical-disease relationships for three uncurated chemicals. Results showed that ANNs have the potential to predict disease associations for uncurated chemicals and to guide research for curated chemicals that may require further toxicological testing.

To my wife and grandpa

Acknowledgments

I would like to thank my advisor Dr. Willie Harper for his guidance and direction throughout the process of this thesis effort. His input was extremely valuable and greatly appreciated. I would also like to thank committee members Lt Col Tay Johannes and Lt Col Dirk Yamamoto for their support and insight. Additionally, I would like to thank Dr. Michael Grimalia for his efforts in obtaining data files from the CTD.

Table of Contents

	Page
Abstract	iii
Acknowledgments.....	v
Table of Contents	vi
List of Figures	viii
List of Tables	ix
I. Introduction	1
Significance of Research.....	4
Implications of Research.....	5
II. Literature Review	6
The relationships between chemical properties and disease	6
Using Artificial Neural Networks in medical research	7
<i>ANN in diagnosis</i>	9
<i>ANN in predictions</i>	11
III. Research Objectives.....	13
IV. Methodology.....	15
Overview	15
Chemical Classification System.....	15
Data	17
<i>Chemicals</i>	18
<i>Diseases</i>	31
Artificial Neural Network	33
Input and Output Data	34
Simulations.....	36
Analysis and Results	38
V. Analysis and Results	40
ANN Training Results	40
ANN Model Performance for Curated Chemicals	56
<i>Effect of Training Ratio on Model-Predicted Disease</i>	56
<i>Chemical Trends for Undertrained Model Simulations</i>	58

<i>Chemical Trends for Overtrained Model Simulations</i>	61
ANN Model Performance for Uncurated Chemicals	63
VI. Conclusions and Recommendations	68
Research Conclusions	68
Recommendations for Future Research	70
Appendix A: MATLAB ANN Code.....	71
Appendix B: Input and Output Matrices.....	75
Appendix C: Additional MATLAB Training Sessions	106
Appendix D: TVT Graphs	122
Appendix E: List of Training Functions	128
Appendix F: Training Function Plots	130
Appendix G: Uncurated chemical data	146
References	148
Vita	154

List of Figures

	Page
Figure 1: New Chemical Classification Example.....	17
Figure 2: Artificial Neural Network Example	34
Figure 3: Typical MATLAB ANN Training Session	41
Figure 4: Typical MATLAB Actual Disease versus ANN Derived Disease Outputs Plot	44
Figure 5: Typical MATLAB ANN Performance Plot	45
Figure 6: Typical MATLAB ANN Mean Squared Error Plot	46
Figure 7: Typical MATLAB ANN Regression Plot	47
Figure 8: Typical MATLAB ANN Training States	48
Figure 9: Typical MATLAB ANN Error Histogram Plot	49
Figure 10: TVT Ratio Effect on the Coefficient of Determination	52
Figure 11: Training Function Effect on the Coefficient of Determination	55
Figure 12: Effect of TVT Ratio on MATLAB ANN Derived Disease.....	57
Figure 13: Effect of Undertrained TVT Ratio (50-25-25 %) on MATLAB ANN Derived Disease	59
Figure 14: Effect of Overtrained TVT Ratio (90-5-5 %) on MATLAB ANN Derived Disease	62
Figure 15: MATLAB ANN Derived Diseases for NCC Cystaphos	64
Figure 16: MATLAB ANN Derived Diseases for NCC 6-HO-BDE-47	66
Figure 17: MATLAB ANN Derived Diseases for NCC 4,4'-diiodobiphenyl	67

List of Tables

	Page
Table 1: Chemical Information.....	20
Table 2: Disease Groupings	32
Table 3: Species Table	35
Table 4: Acetone Input and Output Matrices.....	36
Table 5: The Effect of TVT Ratio on Network Training Statistics	51
Table 6: The Effects of Training Functions on Network Training Statistics.....	54

ARTIFICIAL NEURAL NETWORK PREDICTION OF CHEMICAL-DISEASE RELATIONSHIPS USING READILY AVAILABLE CHEMICAL PROPERTIES

I. Introduction

The natural environment is burdened with a broad range of toxic chemicals, including petroleum products, metals, pesticides, pharmaceutical compounds, organic solvents, and numerous other hazardous substances. Most of these chemicals have the potential to cause ecological harm and they also pose significant risks to human health. Toxicological testing has helped reveal the connections between specific chemicals and health risk factors, but experimental testing on indicator species is expensive and time consuming, while testing on humans is illegal and unethical. There is a need to develop a tool that can accelerate the pace at which we learn how chemicals impact disease. Such a tool would allow the benefits of a given chemical to be weighed against the risks to the environment and public health.

Risks to the environment and to public health are governed by the interactions between chemicals, environmental factors, and the genes that modulate important physiological processes, and there are large databases containing information that can be used to advance fundamental understanding. For example, the Comparative Toxicogenomics Database (CTD) is a publicly available research resource that includes curated data describing cross-species chemical–gene/protein interactions and chemical– and gene–disease associations. The CTD contains over 800,000 chemical–gene interactions, more than 12,400,000 gene-disease associations, and over 1,300,000

chemical-disease associations (Davis et al., 2013). These data can be used to develop insights into complex chemical–gene and protein interaction networks.

The existing CTD data can be used to develop a model that can predict the effect of chemical structure on public health risk. For such a model, Artificial Neural Networks (ANN) can be used. ANNs are flexible mathematical models that are capable of identifying complex nonlinear relationships between input and output data sets. These models are especially useful when it is too difficult to use conventional mathematical equations. ANNs recognize patterns and they work by converting input data into numerical values that are propagated through a network of neurons. The network of neurons processes the data given to the network by attempting to find patterns in the data so that inputs can be correlated outputs. ANNs have been used for a wide range of environmental and public health applications and they are ideal when there is a large amount of data available for ANN development.

One obstacle in investigating chemical-disease relationships is the lack of a useful chemical classification system; one that uses specific chemical characteristics to assign chemical identification numbers. Currently, several individual classification systems provide unique classification numbers for each chemical; however, these numbers are not related to the properties of the chemical and often are randomly assigned. Therefore, developing a modified chemical classification system is an important task for the development of chemical–disease relationships. It would permit policy makers and scientists to anticipate diseases that would be likely associated with new chemicals or existing chemicals that require further testing.

The Environmental Protection Agency (EPA) and National Science Foundation (NSF) have expressed interest in chemical-disease relationships for the purpose of characterizing chemical lifecycles. In 2013, as part of a joint solicitation, the EPA and NSF requested research be conducted that studies the lifecycles of synthetic chemicals, including a focus on their impacts on human health and the ecology (National Science Foundation, 2013). ANNs could provide an appropriate tool to investigate chemical lifecycles, especially when analyzing chemical-disease associations. Understanding how chemicals interact in a given environment and how they could affect surroundings play a role in the lifecycle of a chemical.

The EPA could also use an ANN tool to add important chemical association information to the Toxic Substances Control Act (TSCA) inventory. When the TSCA was implemented in 1976, over 62,000 chemicals were grandfathered into the inventory without any knowledge of their potential affects (Environmental Protection Agency, 2013). In the past 38 years, the number of chemicals in the TSCA inventory has grown to over 84,000, yet only 4 chemicals are specifically addressed within the TSCA document and only a few others have been regulated or banned in the United States (Congressional Digest, 2010). With so many chemicals with unknown chemical-diseases associations existing in the TSCA inventory, a simple analytical tool to generate chemical-disease association predictions may provide valuable information for potentially unknown harmful chemicals. Using a predictive ANN to generate potential chemical impacts could increase the usefulness of the TSCA.

Significance of Research

With the ability to generate predictions of unknown chemical-disease relationships, the ANN provides the possibility of being a useful tool for researchers in the science and medical fields investigating the potential effects of new chemicals. A network that could point researchers towards the effects a chemical will have could help save valuable time and resources when it comes to creating and testing chemicals. When used as a screening and prioritization tool, an ANN may be useful in influencing where researchers begin testing and analyzing chemicals. As the network is expanded through future research, it could potentially be used to predict potential interactions other than just chemical-disease associations. Refining the classification number and training the network with different outputs could allow the network to predict how chemicals may interact if released in to a natural environment. An ANN with this type of capability could be adjusted to work with the Environmental Protection Agency and National Science Foundation's research request for using networks to characterize the lifecycle of chemicals. Additionally, valuable information could be added to the TSCA inventory providing data on potentially harmful chemicals which were grandfathered into the system with no known associations. The true significance of using an ANN to predict chemical-disease associations will become more evident as further research and testing is done to refine the ANN model. As the model becomes more efficient and produces more accurate results, it will be more useful to the scientific community in the screening of new chemicals.

Implications of Research

After investigating the research objectives and analyzing the results of the ANN simulations, it can reasonably be assumed that a MATLAB ANN can be used to analyze chemical and disease data and formulate a network that can possibly predict future chemical-disease associations. The creation and use of a new chemical classification system with an ANN was also demonstrated and results show that a new classification method could be advantageous when working with chemical-disease interactions. Although the classification system developed worked for the simulations conducted in this research, it does not mean that the classification represents the best method or uses the most appropriate chemical properties in the classification number. However, it does indicate that a classification number based on chemical attributes is certainly a possibility and be useful in research and experimentation. Similar to the classification system, the ANN shows the potential for developing networks that can predict chemical-disease relationships; however, the current network may not provide the best performance possible. Training-validating-testing (TVT) ratios and training functions play important roles in the development of the ANN and show strong correlation to how well the network performs, but there are many other factors that can be edited and tested that could improve network further. Using data from the CTD shows that a network could be created on a larger scale and not be bound to specific groups of chemicals or diseases. Combining the CTD data with the new classification system and ANN confirms that chemical-disease association prediction can be accomplished on a large scale, not just in smaller quantitative structure-activity relationship research studies.

II. Literature Review

The relationships between chemical properties and disease

Several studies have explored the relationship between chemical properties and health risk factors or physiological impacts on indicator species. For example, Schultz et al., 2002 discovered positive relationships among 120 different aromatic compounds and estrogenicity based on the number of hydrogen bond donating groups in the aromatic compound. They also found that the number of hydrogen bond accepting groups was negatively linked to estrogenicity. Fang et al., 2001 looked at 230 natural and synthetic steroids and discovered that estrogenicity related negatively to the number of hydrogen bond donating groups in the steroid. They also discovered that estrogenicity was positively linked to the octanol-water partition coefficients of the steroids. Lipinski et al., 2012 expanded on this research looking at 2500 organic compounds and discovered similar results to that of Fang et al. Quantitative structure activity relationships are not limited to only estrogenicity, as they have been used to model numerous other chemical-disease relationships. Ren, 2002 found that hydrophobicity and hydrogen bonds could be used to predict the toxicity of a chemical and Svetnik et al., 2003 determined molecular weight could be used to predict a chemical's biological activity. Wu et al., 2013 also discovered that hydrophobicity and electron density can predict antibacterial qualities of chemicals. This previous work shows that structural properties of chemicals can be used to predict associations between chemicals and other factors.

Using Artificial Neural Networks in medical research

ANNs are flexible, mathematical models capable of identifying complex, nonlinear relationships in data sets. They are capable of discovering patterns in large amounts of data and have been shown to be useful in environmental and public health applications (Beale et al., 2013). ANNs take a set of input and output data and develop correlations between the two data sets by using hidden layers of mathematical formulas, weights, and biases. The formulas are determined by the type of training function specified to be used by the network during the simulation. The weights and biases are placed on the input data as the network is tested and they can be adjusted to help improve network performance. After testing the known input and output data with the training function formulas, weight, and biases, the network derives outputs that are compared to the actual outputs.

When setting up an ANN, two important parameters of a network are the TV ratio and the training function. TVT ratios establish how the data is divided for use in training, validating, and testing the network model. Training functions are the algorithms that determine how the network trains the data while it attempts to find patterns and correlations between the input and output data (Beale et al., 2013). The use of appropriate TVT ratios is important for optimizing network performance because the ratios will determine if a network is undertrained or overtrained. Seguritan et al., 2012 found that testing different training and validation ratios did not provide any significant difference in the overall network performance, but adjusting the testing ratio did show potential for increasing performance. Ahmad and Gromiha, 2003 calculated high ANN

prediction accuracy rates when using TVT ratios that used a majority percentage of the data training the network and Guyon, 1997 found that the ratio of validation data to training data should be between 10 and 25 percent. Singh et al., 2011 compared three training functions in a neural network and found the trainbr function provided the best network performance, but that more than three functions should be tested to truly find the function that best fits the network. Guenther and Frauke, 2010 showed that resilient backpropagation functions performed well in regression ANNs but indicated only three types of functions were tested and other functions may provide similar or better results. Ferrari and Stengel, 2005 found that algebraic training functions may be used to create linear correlations from non-linear datasets with multiple input and output variables.

Overall, research has shown that ANN can be useful in diagnostic and predictive applications when provided the proper data. ANNs have been used in the medical community to address concerns related to specific diseases or groups of diseases. For example, Stephan et al., 2009 used ANNs to distinguish between benign and malignant prostate cancer and Santos-Garcia et al., 2004 used ANNs to predict morbidity from cardio respiratory failure as a result from non-small cell lung cancer pulmonary resection. Curtis et al., 2001 used ANNs to associate genotypes with common human diseases and Sheppard et al., 1999 used ANNs to predict the risks of contracting cytomegalovirus disease after kidney transplantations. Nguyen et al., 2002 used ANNs to predict patient susceptibility to meningitis.

Nearly all of the data used in the ANN analyses for clinical and medical research comes from hard to obtain data or data that requires a great deal of effort to acquire. This

hard to obtain data often requires additional testing and information gathering to acquire the data needed for the ANN. This often requires a great deal of time and resources from the medical personnel. For example, Song et al., 2005 used ultrasound image results and interpretations from physicians to investigate the ANN diagnosis of breast masses and Viazzi et al., 2006 had to obtain cardiac and vascular ultrasound information from physicians and adjust it to work in the ANN model. While useful in medical the medical field, many ANN applications require additional data or testing to successfully use the network.

ANN in diagnosis

ANNs have shown potential to be used in helping doctors diagnose lung diseases by analyzing clinical and radiological factors in addition to relying on chest radiographs. Abe et al., 2002 and Abe et al., 2004 presented evidence that radiologists could use ANN output data, in addition to x-rays, to diagnose lung diseases. Their findings indicated that using clinical factors in an ANN could potentially prove to be more useful when diagnosing interstitial lung disease. Ashizawa et al., 1999 also found that ANNs used by radiologists increased the accuracy of lung diseases diagnosis. Research performed by Feng et al., 2012 discovered that ANN proved capable of diagnosing lung cancer as well differentiating it from benign lung disease, gastrointestinal cancers, and control patients by analyzing various blood levels in patients.

ANNs are not only limited to be used in diagnosing lung disease. In 2009, Babaoğlu et al. concluded that ANN could be used to analyze exercise stress testing data

to correctly diagnose coronary artery disease as well predict the locations of lesions near the heart. Lux et al., 2013 found that hereditary hemorrhagic telangiectasia could be diagnosed by obtaining mid-infrared spectroscopy from blood plasma and analyzing the data through an ANN instead of conducting the typical and costly clinical tests. Matsuki et al., 2002 found that by taking clinical parameters and radiologic findings from high-resolution CT scans and analyzing the data with an ANN, that radiologists could accurately diagnose nodules as benign or malignant without having to conduct further invasive testing on the patients. Arterial blood gas values were predicted based off of venous blood gas values in an effort to better assess patients with acute exacerbations of chronic obstructive pulmonary disease (AECOPD) by Raoufy et al., 2011. Arterial blood gas values provide the best diagnostic evidence of AECOPD but can be difficult to obtain. Using an ANN to correlate venous blood gas values to arterial blood gas values provided an accurate method to detect AECOPD hypercarbia. Deng et al., 1999 found that when ANNs were combined with MRIs, physicians were able to successfully diagnose Alzheimer's disease in potential patients and Mataka et al., 2006 found that the accuracy of radiologist diagnosis of hepatic masses increased when ANN were used to analyze nine clinical parameters from computed tomographic scans. Hamilton et al., 2006 successfully used ANNs to discriminate between parkinsonian syndrome and essential tremors based on the ratio of tracer accumulation between the caudate nucleus and putamen. This model provided a tool to diagnose parkinsonian syndrome early without confusing it with essential tremors.

ANN in predictions

In addition to aiding medical professionals in diagnostics, ANNs have also been shown to be useful in predicting diseases as well. The difference between predicting and diagnosing diseases is diagnosing is identifying a disease or condition that is already present. Predicting uses current data to forecast potential future disease or conditions without any current symptoms. For example, Biglarian et al., 2012 found that ANNs were useful in predicting distant metastasis in colorectal patients and that the ANN models were more accurate than logical regressions models. Colak et al., 2008 created an ANN model that highlighted promising results for predicting coronary artery disease without the need to invasive testing for diagnosis. The work of Cucchetti et al., 2007 demonstrated that ANNs were more accurate than the current model for end-stage liver disease used to prioritize patients with liver cirrhosis for donor organs and Ghoshal et al., 2008 used ANN to predict the mortality of patients with cirrhosis of the liver. Using ANNs for this could help doctors better prioritize transplant candidates, potentially reducing the mortality rate of patients waiting for donor organs. Dagli et al., 2012 used an ANN to predict anemia in patients with Behcet disease. Their model provided a 99% correct anemia prediction rate using prohepcidin and hepcidin levels as well as several other common blood parameters. El-Solh et al., 1999 found that ANNs were more accurate in predicting pulmonary tuberculosis than medical assessments performed by physicians. Recurrence of non-invasive transitional cell carcinoma of the urinary bladder was predicted in an ANN model by Fujikawa et al., 2003 and the model proved to be more accurate than current prognosis techniques. Matsui et al., 2002 developed an ANN

model that showed promising results for being able to replace old methods of predicting prostate cancer in Japanese men. Although in need of further refinement, the model could be used to predict the pathological stage of prostate cancer. Ning et al., 2006 predicted levels of hypertension using physician and patients comment data in an ANN. ANNs were shown to be a possible tool for predicting Graves' disease in patients as a result of antithyroid drug withdrawal by Orunesu et al., 2004 and Salvi et al., 2002 found that ANN could be used to predict the progression of thyroid-associated ophthalmopathy.

The articles discussed in this literature review represent a sample of the articles that can be found on these topics. It is meant to provide an understanding of previous work conducted with chemical-disease relationships and ANNs in the scientific and medical research communities by showcasing relevant research. From analyzing previous research related to chemical-disease relationships and the use of ANNs in investigating chemical-disease associations, it can be concluded that with proper set-up, ANNs can be used to predict potential disease associations from various variable inputs. These previous efforts help to identify and define the research objectives for optimizing ANN performance and then using the network to predict chemical-disease associations.

III. Research Objectives

From the analysis of the literature review, several areas of research became apparent that would attempt to solve the issues presented in Chapter 1. The research objectives of this thesis are as follows:

- 1) What chemical structural characteristics should be used to create a new chemical classification system capable of being used in identifying chemical-disease associations? The first problem addressed will be the creation of a new chemical classification system. Reviews of previous research efforts will be used to highlight potential chemical properties that could be successfully used to create a new numbering system.
- 2) How can MATLAB ANN capabilities be used to connect chemicals to diseases using chemicals structural properties? Before a predictive ANN can be created in MATLAB, it needs to be shown that the MATLAB ANN can properly analyze the input and output data of the new chemical classification system obtained from the CTD. Showing that the MATLAB ANN can be used for this purpose establishes the foundation needed to continue on to the next phase of research.
- 3) What TVT ratio provides the best network performance? Investigating the best TVT ratio to use with the chemical and disease data in the network is

important so that the network performs at the highest level possible. If a suboptimal TVT ratio is used, the predictability of the network will suffer.

- 4) What training function provides the best network performance? A proper training function used in the network is equally as important as a proper TVT ratio. The training function establishes the algorithm the network will use to train the data and update the weights and biases of the network. There are several training functions to choose from in the MATLAB ANN and each one trains the data differently. Determining the proper training ratio increases the prediction potential of the network.
- 5) How can the network be used to predict diseases that are linked to uncurated chemicals? The final step in addressing the problem statement is determining if the network can be used to accurately predict disease outputs. This will be done by creating an ANN with known chemical-disease relationship using curated chemical data from the CTD. Then, uncurated chemicals will be entered into the system and the outputs will be analyzed to see if the model can generate valid predictions.

IV. Methodology

Overview

This chapter will discuss the methods used to acquire, test, and analyze the data in attempting to show that a predictive ANN model can be created to correlate chemicals to unknown disease associations. The data source and required data information will be explained as well as how the ANN will be used to analyze the data. The three simulations needed to complete the research will be addressed as well as how the output data will be analyzed to obtain the final results.

Chemical Classification System

A new numbering system was created that attempted to incorporate specific qualities of the chemical into the classification number while still ensuring that each chemical would have an individual and unique number. From investigating previous research of quantitative structure activity relationships, several studies demonstrated that readily available physical properties of chemicals could be used to predict a chemical's effect when used in an appropriate model (Schultz et al., 2002, Fang et al., 2001, Lipinski et al., 2012, Ren, 2002, Svetnik et al., 2003). Using readily available chemical properties allows for a chemical classification system number to be created without extensive testing or data collecting.

Three chemical traits that proved useful, particularly in a chemicals effect on estrogenicity, were molecular weight, hydrogen donors, and hydrogen acceptors (Lipinski

et al., 2012). In addition to being successfully used in previous research, these three chemical traits are also fairly simple to obtain. Without performing complex testing, the molecular weight can be easily calculated from the chemical formula while the hydrogen acceptors and donors are added up based on the number of lone pair electrons and atoms bonded to at least one hydrogen atom. For example, the oxygen atom in water has one free pair of electrons so water has one hydrogen acceptor. The oxygen atom in water is also connected to two hydrogen atoms so water has one hydrogen donor.

The numbering system created gave each chemical a ten-digit number that was exclusive to that specific chemical. The first six digits of the number correspond to the molecular weight of the chemical including two decimal places. The seventh and eighth digits represent the number of hydrogen acceptors and the ninth and tenth digits represent the number of hydrogen donors. Figure 1 shows an example of how the new chemical classification number is created for water. Water has a molecular weight of 18.015 g/mol, 1 hydrogen acceptor, and 1 hydrogen donor. Inputting these numbers into the new classification system, the new number generated for water is 0018020101.

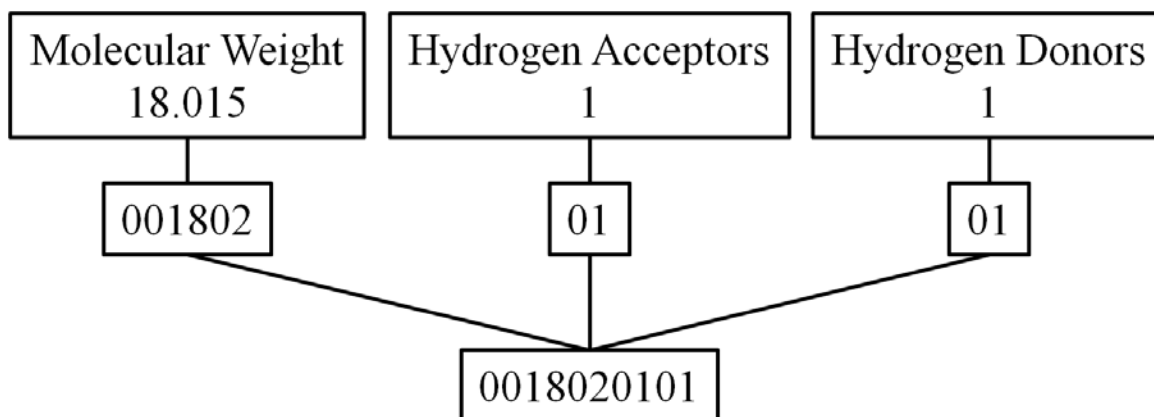


Figure 1: New Chemical Classification Example

The molecular weight in the classification number is assigned to the hundredths place to provide sufficient accuracy and uniqueness for identifying the chemical. The number of hydrogen acceptors and hydrogen donors further reduce the possibility that two different chemicals would have the same classification number.

Data

The data used in the MATLAB simulations comes from the online CTD. The CTD is developed and maintained through a joint effort between North Carolina State University and Mount Desert Biological Laboratory and also receives financial support from the National Institute of Environmental Health Sciences. The primary goal of the CTD is to advance the understanding of the effects of environmental chemicals on human health through studying the relationships between chemicals, genes, and diseases. This online database is a collection of curated and uncurated data containing information for chemicals, genes, and diseases. Curated data is data that has peer-reviewed, scientific research to prove existence of the data relationship (i.e. chemical-disease association).

Uncurated data does not have any literature showing interactions with other factors. The CTD contains over 800,000 chemical–gene associations, 12,400,000 gene-disease associations, and 1,300,000 chemical-disease associations with additional data and updates made weekly. Chemicals, genes, and diseases that are curated have been organized based on documented scientific research. Uncurated chemicals, genes, and diseases do not have the support of peer-reviewed research. While uncurated data do not have documentation to prove associations, the CTD does list possible associations through inferences. An inferred association between a chemical and a disease is established with a curated chemical-gene and gene-disease relationship. For example, acetone has a curated relationship with the catalase (CAT) gene and the CAT gene has been shown to affect asthma in humans. There is no direct link between acetone and asthma but it can be inferred through the curated relationships with the CAT gene (Davis et al., 2013)

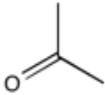
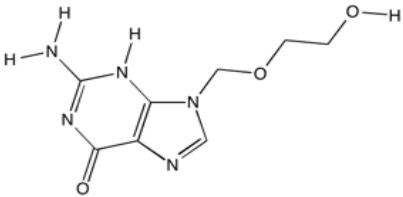
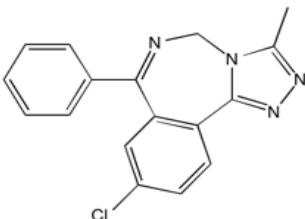
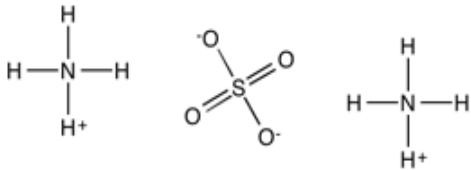
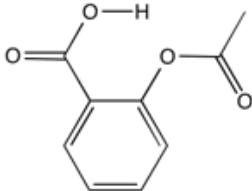
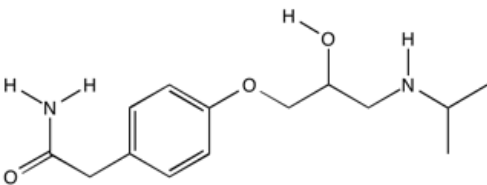
The CTD website offers several organization and research functions that can be used to explore how the chemicals, genes, and diseases interact and related to one another. These functions allow users to research specific categories within the data and specify particular relationships of interest. In addition to the search functions, user can download entire sets of the database to use in simulations and research.

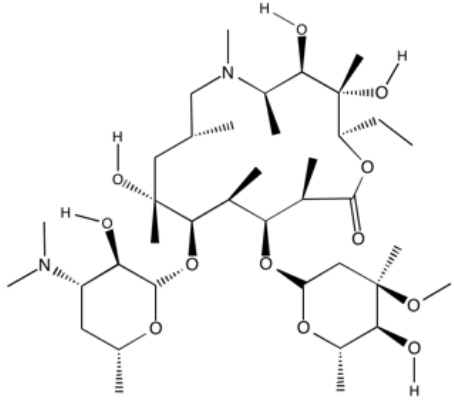

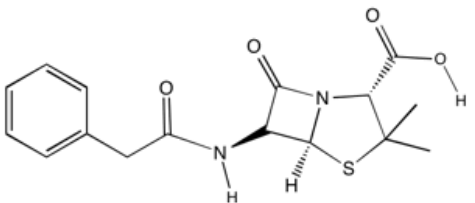
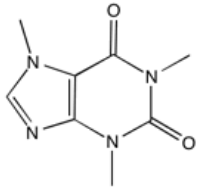
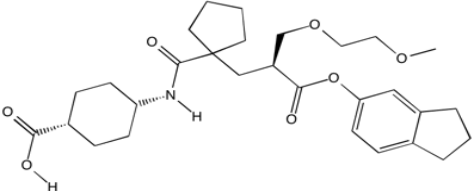
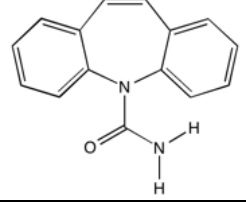
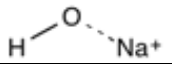
Chemicals

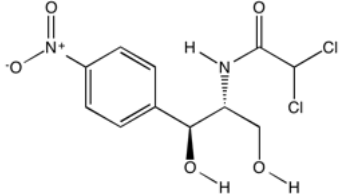
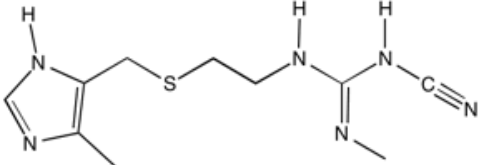
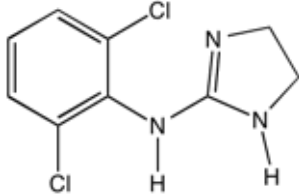
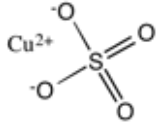
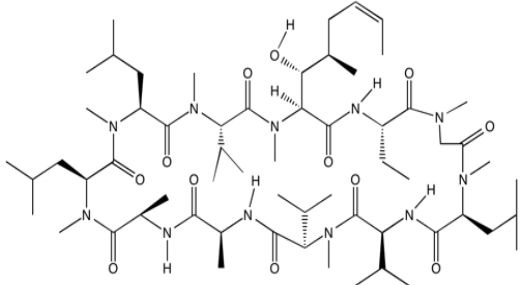
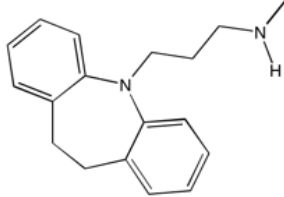
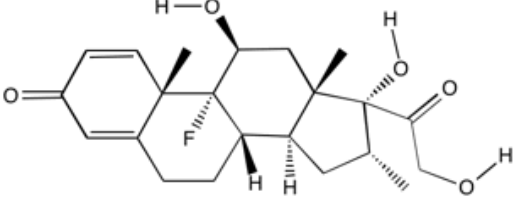
The chemicals used in the MATLAB ANN simulations were randomly chosen from the CTD. As chemicals were selected, they were checked to make sure each

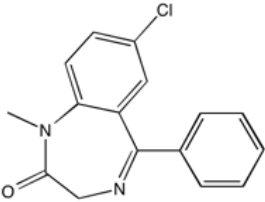
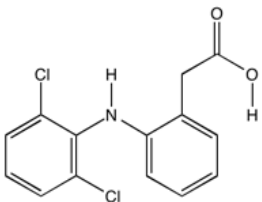
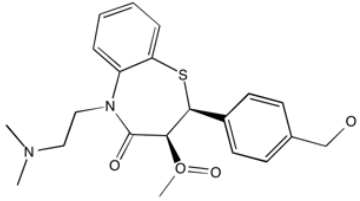
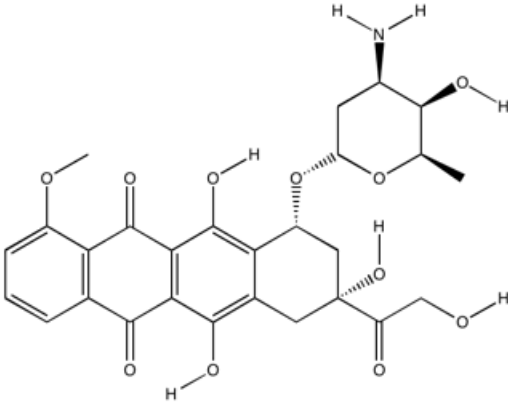
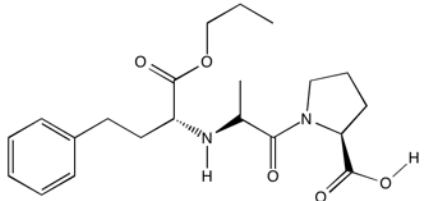
chemical was curated so that known disease associations existed to use in the network. Chemicals were also traced back to their most common ancestor chemical if they were a descendant. Descendant chemicals were traced back to a common ancestor to ensure there was enough data to use in the simulations. Once the ancestor chemicals were determined, the molecular weight, hydrogen acceptors, and hydrogen donors were determined and the new classification system number was generated for each chemical. Table 1 shows the list of chemicals used in the simulations, along with the classification number, chemical formula, and chemical structure diagram. The chemical formulas and structure diagrams were included to show the diversity of the chemicals used.

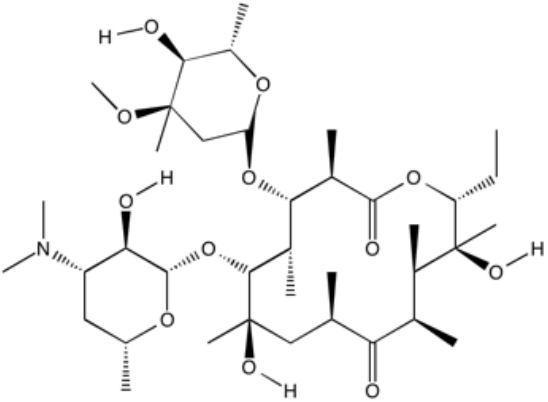

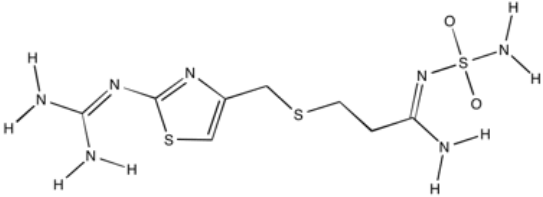
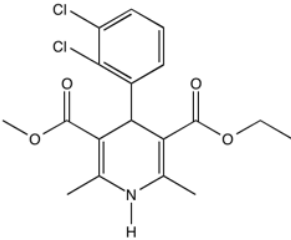
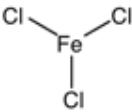
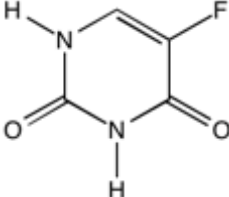
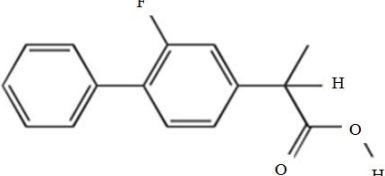
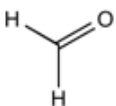
Table 1: Chemical Information

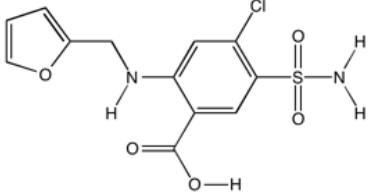
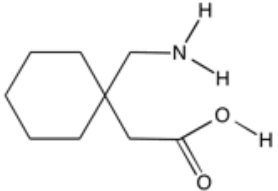
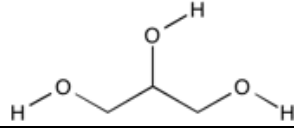
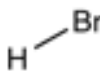
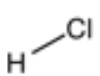
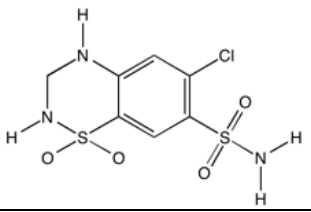
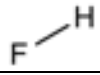
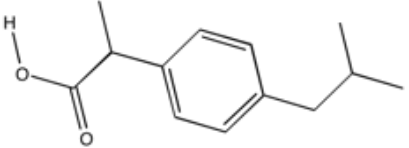
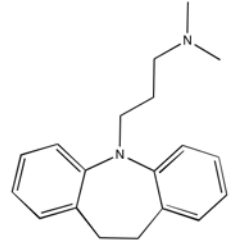
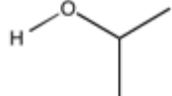
Chemical Name	New Classification Number	Molecular Formula	Chemical Structure
Acetone	0058080100	C_3H_6O	
Aciclovir	0225200804	$C_8H_{11}N_5O_3$	
Alprazolam	0308760300	$C_{17}H_{13}ClN_4$	
Ammonium Sulfate	0132140402	$H_8N_2O_4S$	
Aspirin	0180160401	$C_9H_8O_4$	
Atenolol	0226340403	$C_{14}H_{22}N_2O_3$	

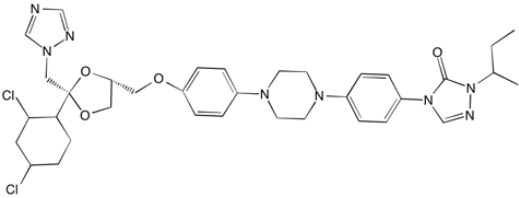
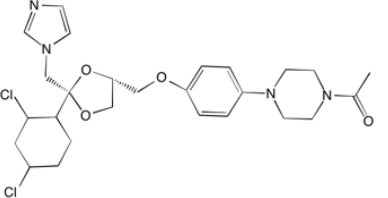
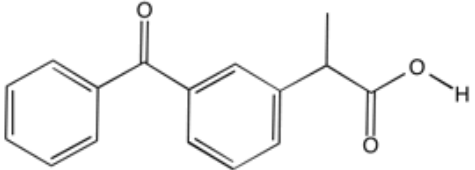
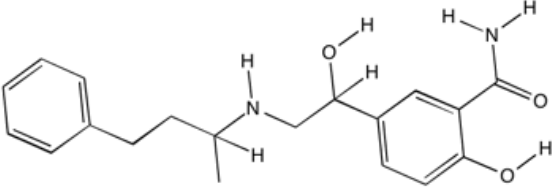
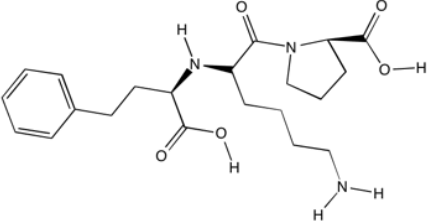
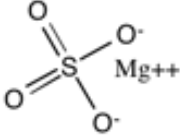
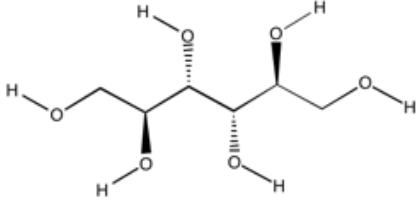
Azithromycin	0748981405	$C_{38}H_{72}N_2O_{12}$	
Benzene	0078120000	C_6H_6	
Benzyl Penicillin	0334390602	$C_{16}H_{18}N_2O_4S$	
Caffeine	0194190300	$C_8H_{10}N_4O_2$	
Candoxatril	0515640702	$C_{29}H_{41}NO_7$	
Carbamazepine	0236270101	$C_{15}H_{12}N_2O$	
Sodium Hydroxide	0039990101	HNaO	

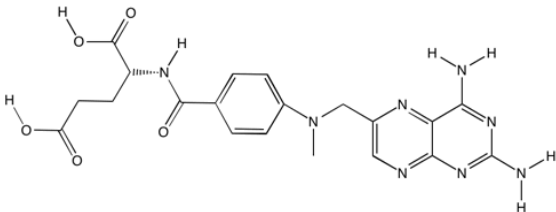
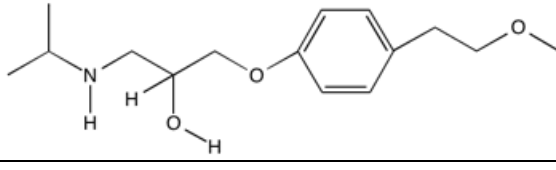
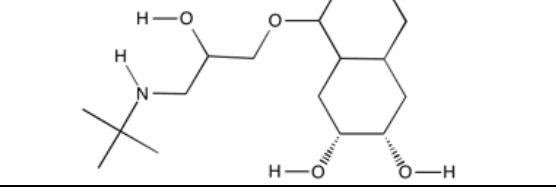
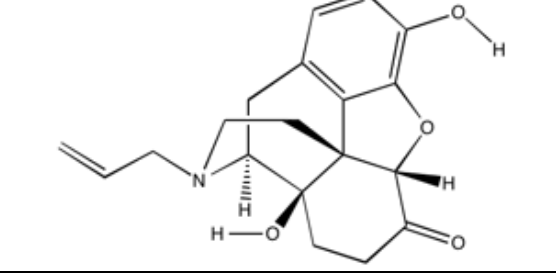
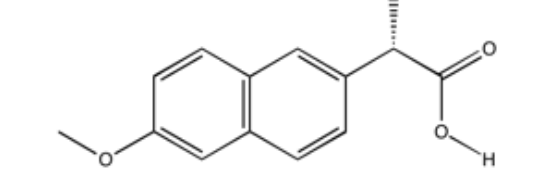
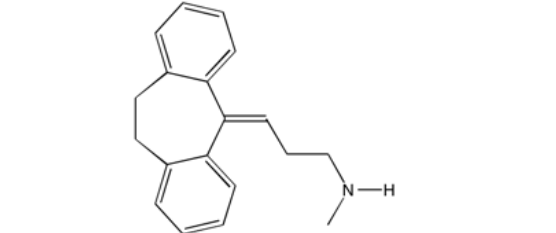
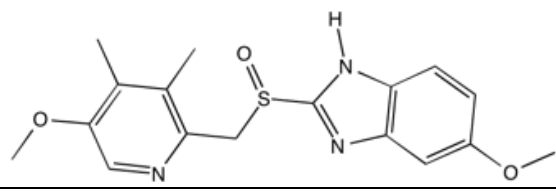
Chloramphenicol	0323130503	$C_{11}H_{12}Cl_2N_2O_5$	
Cimetidine	0252340403	$C_{10}H_{16}N_6S$	
Clonidine	0230090102	$C_9H_9Cl_2N_3$	
Copper Sulfate	0159610400	CuO_4S	
Cyclosporine	1202611205	$C_{62}H_{111}N_{11}O_{12}$	
Desipramine	0266380201	$C_{18}H_{22}N_2$	
Dexamethasone	0392460603	$C_{22}H_{29}FO_5$	

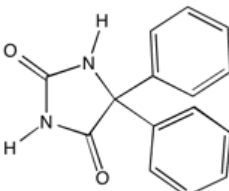
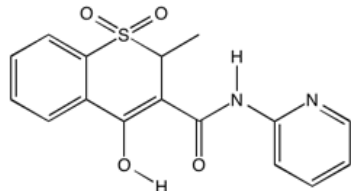
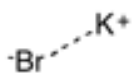
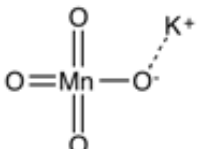
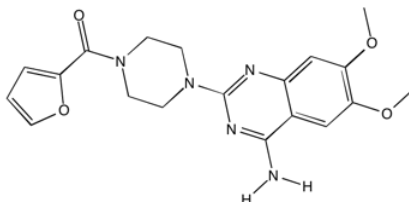
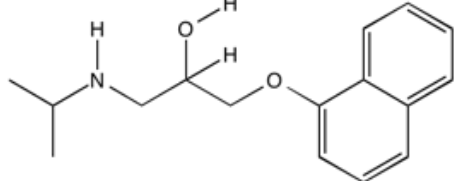
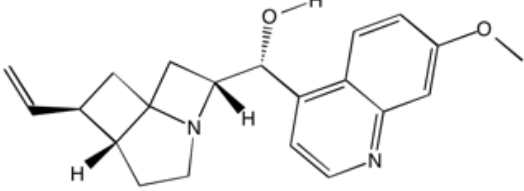
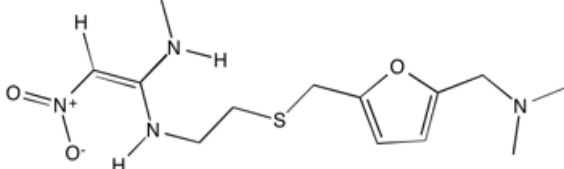
Diazepam	0284740200	$C_{16}H_{13}ClN_2O$	
Diclofenac	0296150302	$C_{14}H_{11}Cl_2NO_2$	
Diltiazem-HCl	0414520600	$C_{22}H_{26}N_2O_4S$	
Doxorubicin	0543521206	$C_{27}H_{29}NO_{11}$	
Enalaprilat	0376450602	$C_{20}H_{28}N_2O_5$	

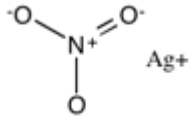
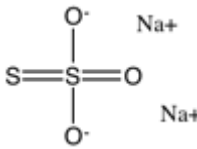
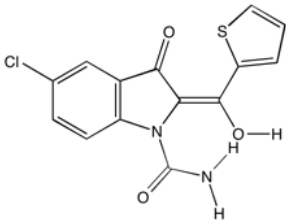
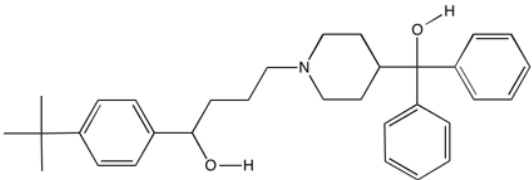
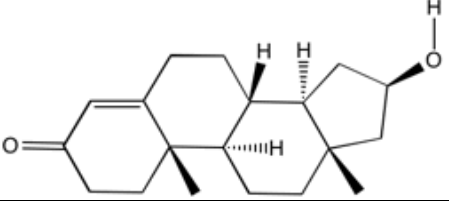
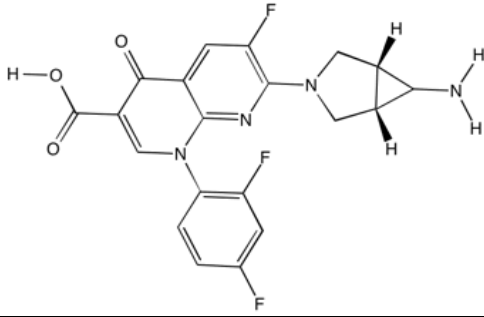
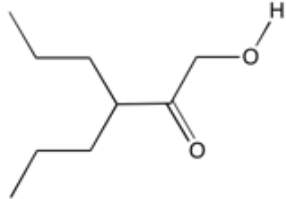
Erythromycin	0733931405	$C_{37}H_{67}NO_{13}$	
Ethylene Glycol	0062070202	$C_2H_6O_2$	
Famotidine	0337450804	$C_8H_{15}N_7O_2S_3$	
Felodipine	0384250501	$C_{18}H_{19}Cl_2NO_4$	
Ferric Chloride	0162200000	Cl_3Fe	
Fluorouracil	0130080302	$C_4H_3FN_2O_2$	
Flurbiprofen	0244260301	$C_{15}H_{13}FO_2$	
Formaldehyde	0030030100	CH_2O	

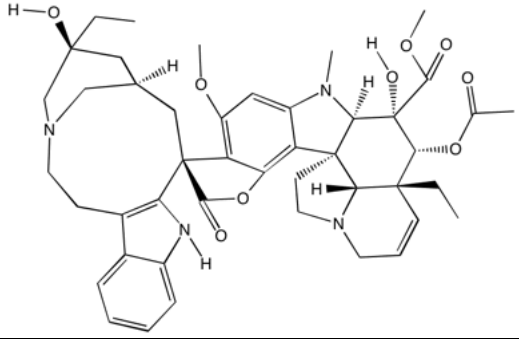
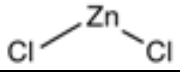
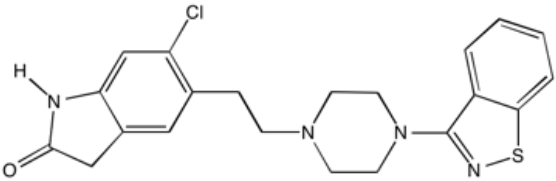
Furosemide	0330740703	$C_{12}H_{11}ClN_2O_5S$	
Gabapentin	0171240302	$C_9H_{17}NO_2$	
Glycerol	0092090303	$C_3H_8O_3$	
Hydrobromic Acid	0080910000	BrH	
Hydrochloric Acid	0036460001	HCl	
Hydrochlorothiazide	0297740703	$C_7H_8ClN_3O_4S_2$	
Hydrofluoric Acid	0020010101	FH	
Ibuprofen	0206280201	$C_{13}H_{16}O_2$	
Imipramine	0280410200	$C_{19}H_{24}N_2$	
Isopropyl Alcohol	0060100101	C_3H_8O	

Itraconazole	0705630900	$C_{35}H_{38}Cl_2N_8O_4$	
Ketoconazole	0531430600	$C_{26}H_{28}Cl_2N_4O_4$	
Ketoprofen	0254280301	$C_{16}H_{14}O_3$	
Labetalol-HCl	0328410404	$C_{19}H_{24}N_2O_3$	
Lisinopril	0405490704	$C_{21}H_{31}N_3O_5$	
Magnesium Sulfate	0120370400	MgO_4S	
Mannitol	0182170606	$C_6H_{14}O_6$	

Methotrexate	0454441205	$C_{20}H_{22}N_8O_5$	
Metoprolol	0267360402	$C_{15}H_{25}NO_3$	
Nadolol	0309400504	$C_{17}H_{27}NO_4$	
Naloxone	0327370502	$C_{19}H_{21}NO_4$	
Naproxen-sodium	0230260301	$C_{14}H_{14}O_3$	
Nortriptylene-HCl	0263380101	$C_{19}H_{21}NO_4$	
Omeprazole	0345420601	$C_{17}H_{19}N_3O_3S$	

Phenytoin	0252270202	$C_{15}H_{12}N_2O_2$	
Piroxicam	0331350702	$C_{15}H_{13}N_3O_4S$	
Potassium Bromide	0119000100	BrK	
Potassium Permanganate	0158030400	MnO_4K	
Prazosin	0383410801	$C_{19}H_{21}N_5O_4$	
Propranolol-HCl	0259350302	$C_{16}H_{21}NO_2$	
Quinidine	0324430401	$C_{20}H_{24}N_2O_2$	
Ranitidine-HCl	0314410702	$C_{13}H_{22}N_4O_3S$	

Silver Nitrate	0169870300	AgNO_3	
Sodium Thiosulfate	0158110400	$\text{Na}_2\text{O}_3\text{S}_2$	
Tenidap	0320760402	$\text{C}_{14}\text{H}_9\text{ClN}_2\text{O}_3\text{S}$	
Terfenadine	0471690302	$\text{C}_{32}\text{H}_{41}\text{NO}_2$	
Testosterone	0288430201	$\text{C}_{19}\text{H}_{28}\text{O}_2$	
Trovafloxacin	0416361002	$\text{C}_{20}\text{H}_{15}\text{F}_3\text{N}_4\text{O}_3$	
Valproic-acid	0144220201	$\text{C}_8\text{H}_{16}\text{O}_2$	

Vinblastine	0810971203	$C_{46}H_{58}N_4O_9$	
Zinc Chloride	0136290000	$ZnCl_2$	
Ziprasidone	0412940501	$C_{21}H_{21}ClN_4OS$	

Diseases

Due to the large number of diseases present in the CTD, the diseases were combined into 27 groups based on the classifications used in the CTD. Rather than associating each chemical with every associated disease, the chemicals were related to the disease group that contained the actual associated disease. Each disease group was assigned a number, 1-27, to represent that disease group in the network. Table 2 shows the 27 diseases groups used in the network and the disease identification number assigned to each group.

Table 2: Disease Groupings

Disease Category	Disease Number
Animal Diseases	1
Bacterial Infections and Mycoses	2
Cardiovascular Diseases	3
Congenital, Hereditary, Neonatal Diseases and Abnormalities	4
Digestive System Diseases	5
Environmental Disorders	6
Endocrine System Diseases	7
Eye Diseases	8
Female Urogenital Diseases and Pregnancy Complications	9
Hemic and Lymphatic Diseases	10
Immune System Diseases	11
Male Urogenital Diseases	12
Mental Disorders	13
Musculoskeletal Diseases	14
Neoplasms	15
Nervous System Diseases	16
Nutritional and Metabolic Diseases	17
Occupational Diseases	18
Otorhinolaryngologic Diseases	19
Parasitic Diseases	20
Pathological Conditions, Signs and Symptoms	21
Respiratory Tract Diseases	22
Skin and Connective Tissue Diseases	23
Stomatognathic Diseases	24
Substance-Related Disorders	25
Virus Diseases	26
Wounds and Injuries	27

Artificial Neural Network

Figure 2 shows an illustration of how the ANN operated for the MATLAB simulations. The ANN took the input and output data for each chemical-disease association and used the training function in the hidden layer to update the weights and biases in an attempt to find patterns and correlations between the input and output data. Based on the pre-determined values in the TVT ratio, the ANN would select the designated amount of data to train the network with. After training, the ANN would then validate the network with the designated amount of data, and finally test the network with the remaining data. In Figure 2, the three output categories (species, dummy variable, and disease) within the ANN represent the actual disease outputs obtained from the CTD. The outputs on the outside of the ANN represent the outputs generated by the ANN during the testing phase of each simulation. The code used in the network simulations can be found in Appendix A.

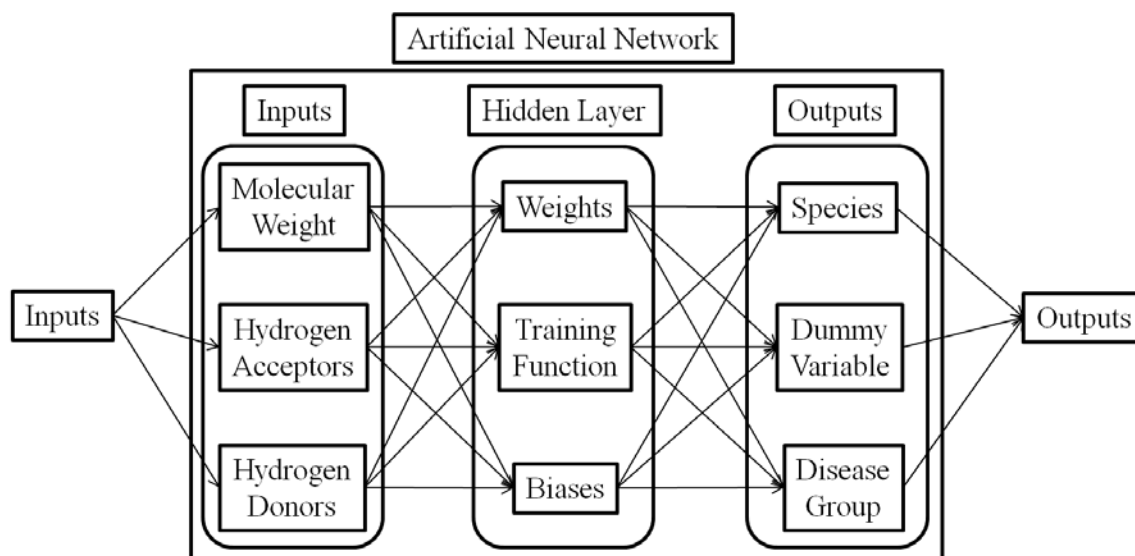


Figure 2: Artificial Neural Network Example

Input and Output Data

Before running the network simulations in MATLAB, the data were first formatted to fit the ANN requirements as defined by the MATLAB user guide. The input and actual associated output data from the CTD were entered into two matrices created in Microsoft Excel from which MATLAB was coded to retrieve the data. The input data were entered into a matrix with three columns, one each for molecular weight, hydrogen acceptors, and hydrogen donors. These three chemical characteristics were left as individual data points for the input data, rather than being entered in one column as the new classification number, to keep the size of the input and output matrices the same and reduce the use of dummy variables. The output data were entered into a matrix the same size as the input data matrix with columns for species, dummy variable, and disease group. Species were given numbers, similar to the disease groups, and the species and

their assigned numbers can be found in Table 3. The species number is specifically related to the species that the chemical-disease association occurs in. For example, acetone has a curated relationship with neoplasms (disease category 15) and this chemical-disease association occurs in humans. A dummy variable was used in the output matrix to keep the matrix the same size as the input data matrix. Zeros were entered for the dummy column values and the dummy column was not used in any of the results analysis.

Table 3: Species Table

Species	Species Number
Humans	1
Dogs	2
Fish	3
Birds	4
Rats/mice	5

The number of rows used for each chemical was determined by the number of disease groups the chemical was associated with. For example, acetone was associated with six diseases groups so it used six rows in both the input and output matrices. The actual input and output data used for acetone can be seen in Table 4. The six rows used in the input matrix all contained the same molecular weight, hydrogen acceptor, and hydrogen donor data. Each of the six rows in the output matrix corresponded to one of the associated disease groups and subsequent related species. The complete input and output matrices can be found in Appendix B.

Table 4: Acetone Input and Output Matrices

Inputs			Outputs		
Molecular Weight	Hydrogen Acceptors	Hydrogen Donors	Species	Dummy	Disease Group
58.08	1	1	1	0	9
58.08	1	1	1	0	12
58.08	1	1	1	0	15
58.08	1	1	1	0	16
58.08	1	1	1	0	17
58.08	1	1	1	0	25

Simulations

Three phases of simulations were used to test the ANN. The first phase of simulation was accomplished to demonstrate that the MATLAB ANN could be used to analyze the chemical and disease data and find appropriate correlations. Initially, 20 chemicals were chosen from the CTD and the input and output matrices were created based on the chemical characteristics and associated species and diseases groups. This simulation used the basic MATLAB ANN code formatted with a TVT ratio of 80-10-10 percent and the default training function. Results from the first phase simulation can be found in Figures 3-9 in Chapter 4.

The second phase of simulations involved testing different TVT ratios and training functions on the network which provided the best network performance. To do this, an additional 55 chemicals were chosen at random from the CTD and the appropriate chemical, species, and disease group data were added to the input and output matrices. First, five different TVT ratios were tested in the network using the default training function: 50-25-25, 60-20-20, 70-15-15, 80-10-10, and 90-5-5 percent. These

five ratios were chosen so that an appropriate TVT ratio could be chosen while also highlighting how the different ratios affect the ANN. Next, each of the 15 training functions was tested in the network using the 70-15-15 percent TVT ratio. The MATLAB training functions used are preprogrammed functions built within MATLAB. All of the functions operated in a feedforward network with backpropagation. A feed forward network is where the data is passed through the hidden layer in a single direction from the input side to the output side. The backpropagation step involves going back to adjust the weights and biases in the hidden layer after comparing the actual outputs to the ANN derived outputs. Each of the training functions are based off of a gradient descent algorithm where the training function attempts to decrease the error in the network by adjusting the weights and biases after each network iteration. The results from the different training ratios indicated that the `trainlm`, Levenberg-Marquardt backpropagation, function generated the best network performance.

The third phase of simulations involved inputting uncurated chemical data into the network that had no known disease associations to see what diseases the network would predict. To accomplish this, a network simulation was first run with the original 75 chemicals using the 70-15-15 percent TVT ratio and `trainlm` training function. The 70-15-15 percent TV ratio and `trainlm` training function provided the best network performance in the second phase of simulations. This established the correlations, weights, and biases for the network to use on the uncurated chemicals. Then the data for three uncurated chemicals were run through the network 10 times and the network-derived disease output results for each trial were recorded. Each uncurated chemical was

input into the network 10 times to allow multiple disease association predictions to occur in the event that a chemical was associated to more than one disease. The network outputs generated had to be rounded to directly correlate them to the whole number designators assigned to each disease group. Several network outputs produced the same disease group more than once for a given input chemical. For duplicate disease predictions, these outputs were consolidated into one output. The derived disease outputs were then compared to research literature to see if the network predictions were correct.

Analysis and Results

Once the simulations were complete in MATLAB, all of the output data was copied into Microsoft Excel for analysis. Analyzing the first phase of simulations was done simply by reviewing the output plots and figures generated by MATLAB at the completion of the simulation. The analyses of the phase two simulations used Microsoft Excel to plot and chart the output data from MATLAB. The primary graph used plotted the actual diseases values versus the ANN derived disease values. When the network produced accurate output results, the plot would follow a linear, one-to-one slope on the graph. Excel was also used to calculate the coefficient of determination (R^2) values to see how well the data followed a linear progression. The analysis also compared the different TVT ratios and training functions based on network parameter values to determine which ratio or function provided the best network performance. Additionally, the effects of undertraining and overtraining the network on ANN derived disease values were taken into account. The third phase of simulations was primarily analyzed by

comparing the ANN derived disease values to research literature in an effort to show that the network had some predictive capability for uncurated chemicals.

V. Analysis and Results

ANN Training Results

The following figures were generated from a typical MATLAB ANN simulation using the default training settings and a TVT ratio of 80-10-10 percent. They show typical results seen during the first phase of simulations when the initial group on 20 chemicals was used in the ANN.

Figure 3 shows a typical MATLAB training session for an ANN. The neural network diagram shows a pictorial of how the network will function, given the requirements established in the network code. At the top of the figure, the diagrams shows the number of inputs and outputs being used and the number of hidden layers. The data division and derivative algorithms are preset within MATLAB and these default settings were used for the various simulations. The training algorithm defines the training function used in the network, which dictates how the data will be trained. The performance algorithm measures how well the network is operating during training. The training and performance algorithms can be changed separately; however, a default performance algorithm will be chosen based on the training algorithm if a specific performance algorithm is not defined.

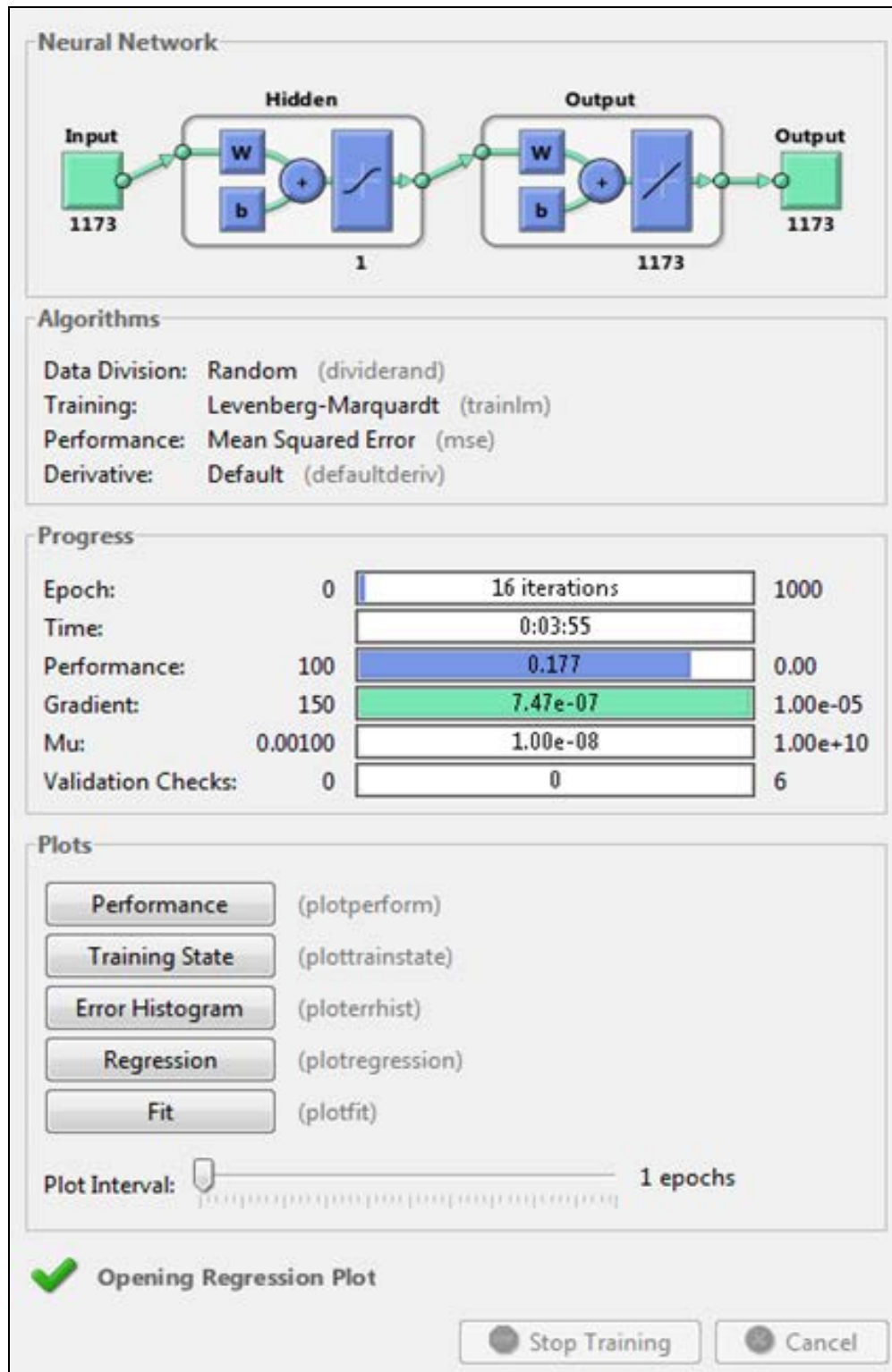


Figure 3: Typical MATLAB ANN Training Session

The progress portion of Figure 3 shows the network performance criteria that provide information on the progress of network training. As the simulation progresses, these performance criteria update to show the current status of the training. The epochs show the number of iterations over the course of the simulation. Performance shows how accurately the network is generating output values compared to the actual output values. The performance values are the mean-squared error of the network so lower performance values indicate higher network training performance. The gradient is an optimization algorithm that measures the adjustments made to the network in relation to the network performance during each epoch. If a network is performing well, smaller adjustments are needed so the gradient will be smaller. Likewise, if a network is performing poorly, larger adjustments are needed to attempt to improve performance so the gradient will be larger. The mu value shows how much the network is required to change the weights and biases placed on the input data to achieve accurate output results. The weights and biases are used to adjust and determine how the input data is used when training the network. If the network finds some input data provides better performance than other data, it will place more weight on the data that increases performance. When a network has high mu values during training, the network is struggling to find weight and bias values that work for the data set. The validation checks show the number of iterations where the simulation's validation performance does not decrease. After the network is trained, it is validated to ensure the training was successful.

All of these performance criteria have upper and lower boundaries and when the appropriate boundary of one of the criteria is reached, the simulation is terminated.

Performance and gradient will both terminate the simulation when they reach their lower limit values. If performance or gradient terminate the simulation, this generally indicates a successful training because lower performance and gradient values correlate to higher network performance. Epochs, mu and validation checks will terminate the simulation when they reach their upper limits. When the number of epochs terminates the simulation, the network has used all of the allotted iterations before the performance or gradient have reached an acceptable level. Network termination due to a high mu value indicates that the network failed to find appropriate weight and bias values. Terminating for a high number of validation checks indicates the network did not train well because it was unable to be validated. Time is the only criterion that does not terminate a simulation as it is simply meant to track how long the simulation takes. Once the network training is complete, there are several figures that MATLAB can create which provide additional information about the network simulation. Figures 4-9 show the different plots that can be generated by MATLAB after the network simulations. Additional training sessions and supporting figures can be found in Appendix C.

Figure 4 shows a typical MATLAB-generated plot of the actual disease numbers plotted against the ANN derived disease. When the plot shows a straight, positive, one to one slope, the ANN derived disease number is close to the actual disease number. The straight, one to one slope seen in the figure suggests that the network was able to correctly derive the appropriate disease number for each chemical.

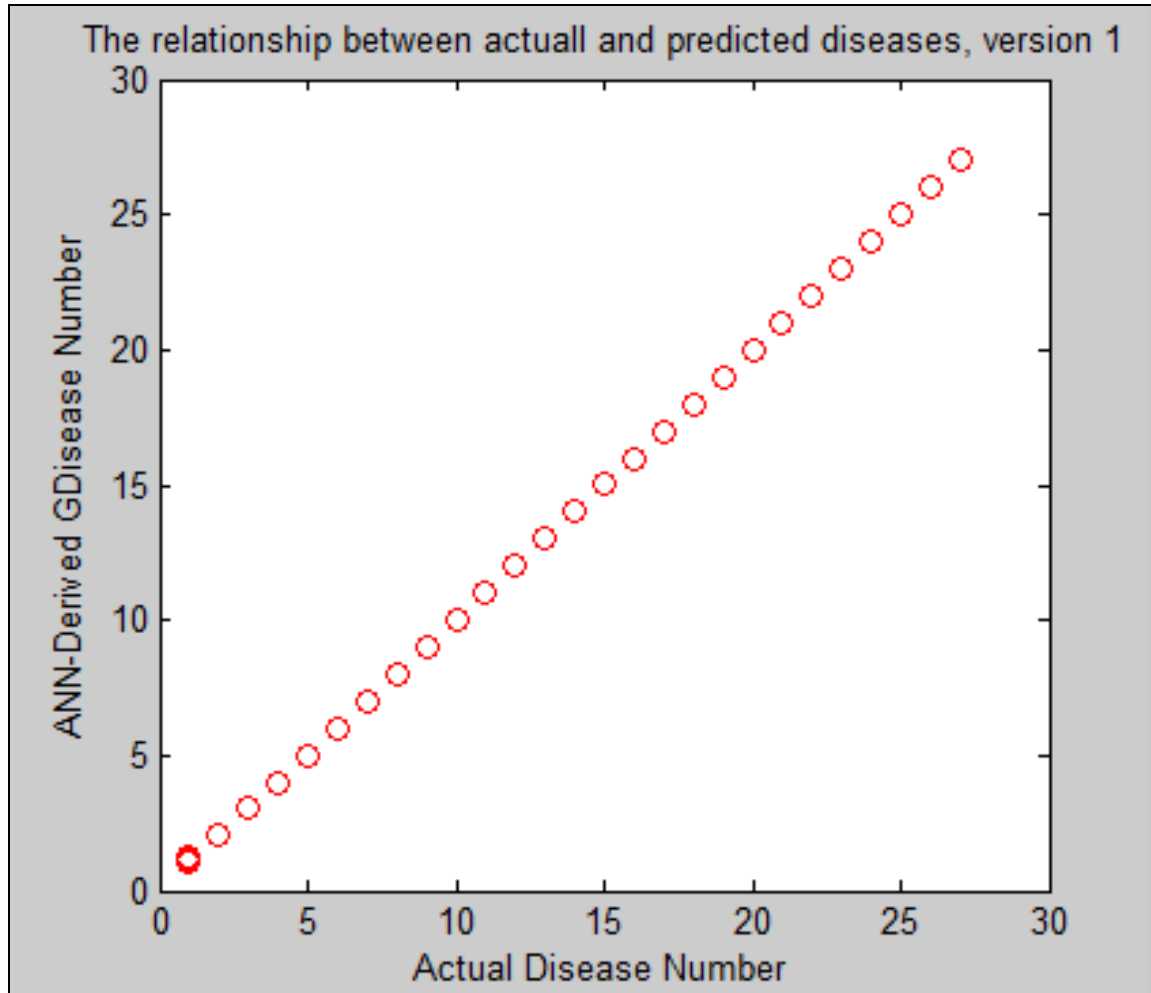


Figure 4: Typical MATLAB Actual Disease versus ANN Derived Disease Outputs Plot

Figures 5 and 6 show the performance (i.e. mean squared error) of the ANN simulation discussed from Figure 3. The two plots look identical because the network used the mean squared error to judge the performance of the network as it was trained. The lower the performance value, or mean squared error value, the better the network is performing because there is less variation between the actual disease groups numbers and the ANN derived disease group numbers. The figures show the performance of the network for each epoch during the simulation. During the simulation, the training

performance decreases until epoch 11 where the performance values plateau. In Figure 5, the validation and test curves follow the same path as the train performance curve.

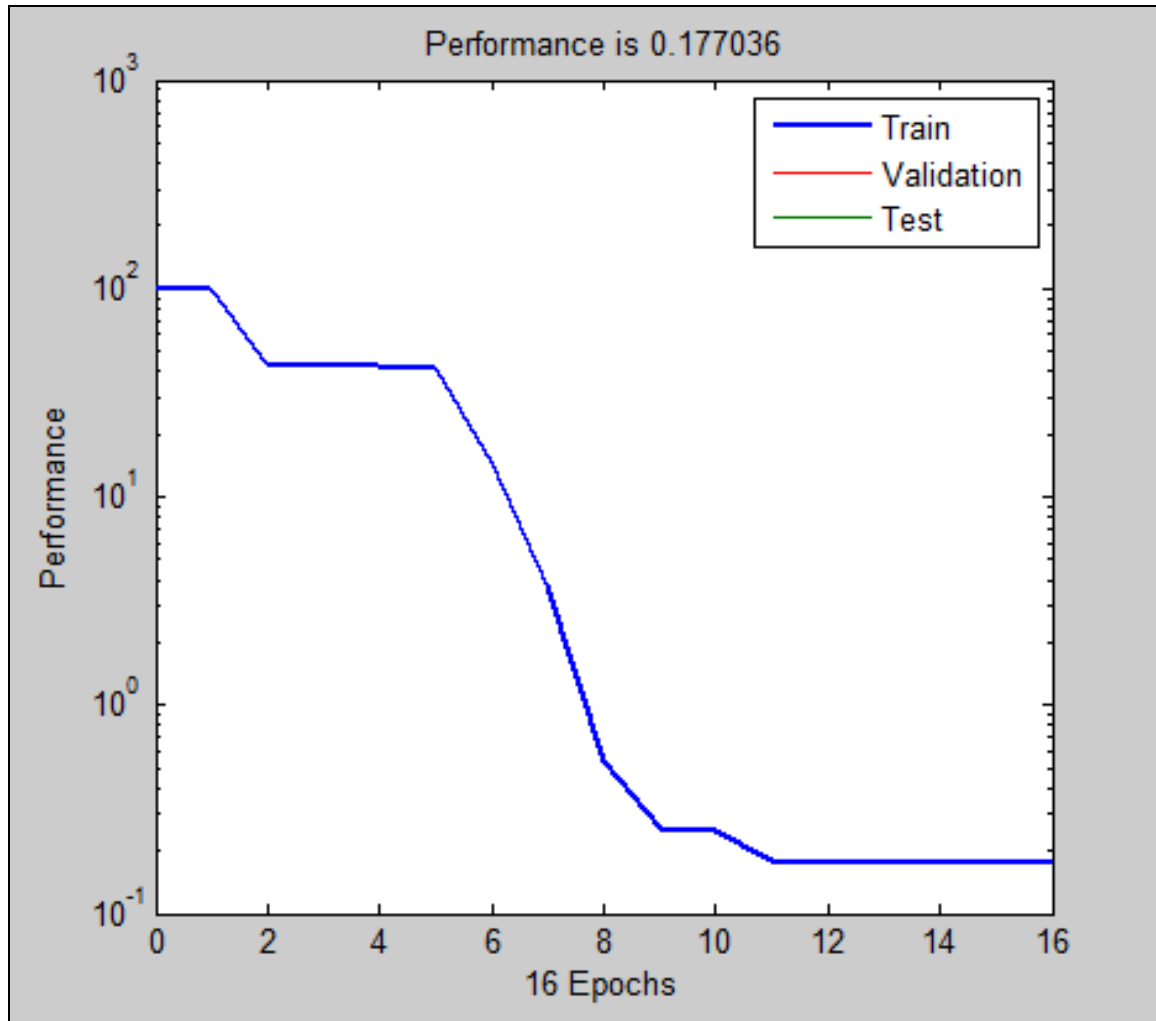


Figure 5: Typical MATLAB ANN Performance Plot

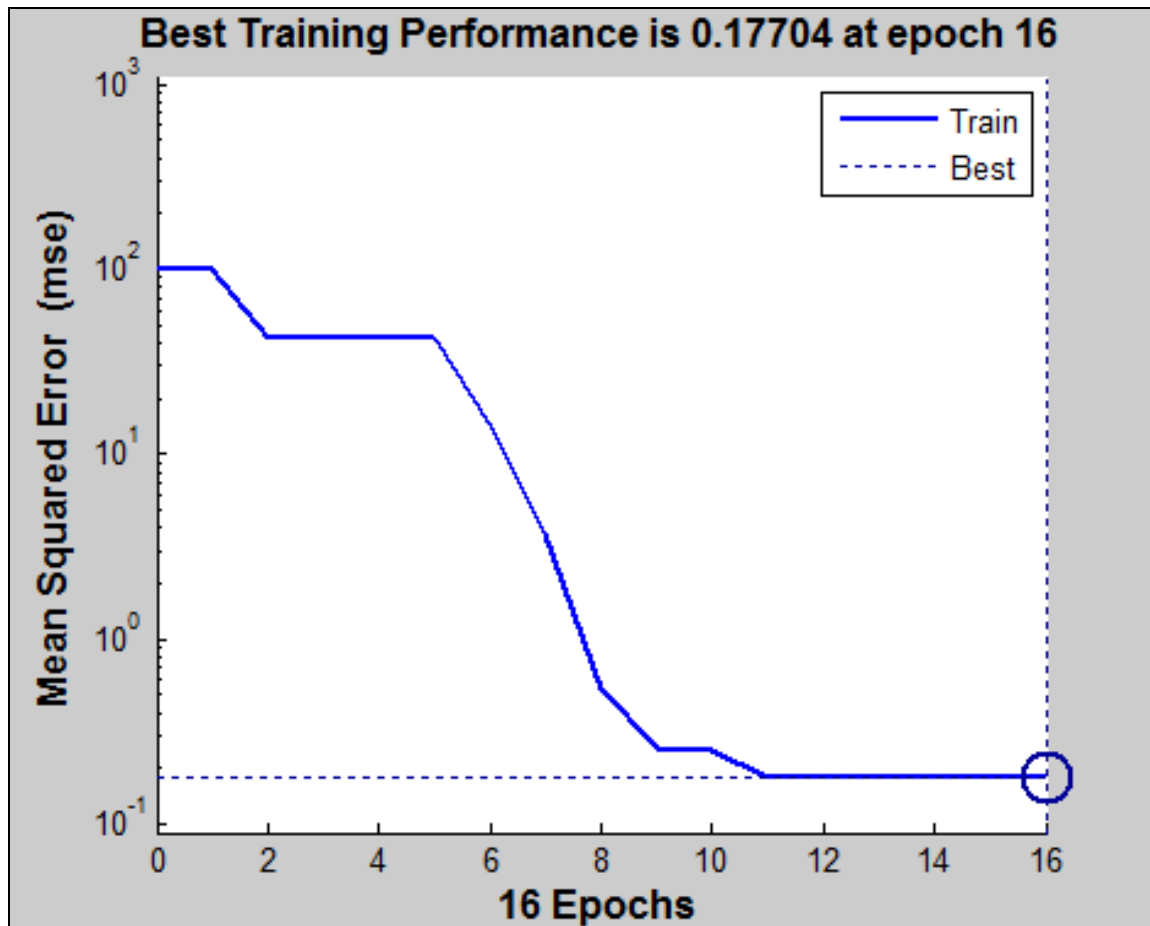


Figure 6: Typical MATLAB ANN Mean Squared Error Plot

Figure 7 shows a typical regression plot produced from an ANN network simulation. Similar to Figure 4, a positive, straight, one to one slope is desired to show that the ANN derived disease numbers match the actual disease numbers. In this plot, the target, x-axis, values represent the actual disease numbers and the output, y-axis, values represent the ANN derived disease numbers. The training R-value shows how closely the derived numbers compare to the actual numbers. An R-value of one would indicate a perfect match of the derived to the actual disease number so the higher the R-value, the

more accurate the Ann derived disease numbers are. The high R-value of 0.9985 for this simulation corresponds to the straight, one to one slope seen in the graph.

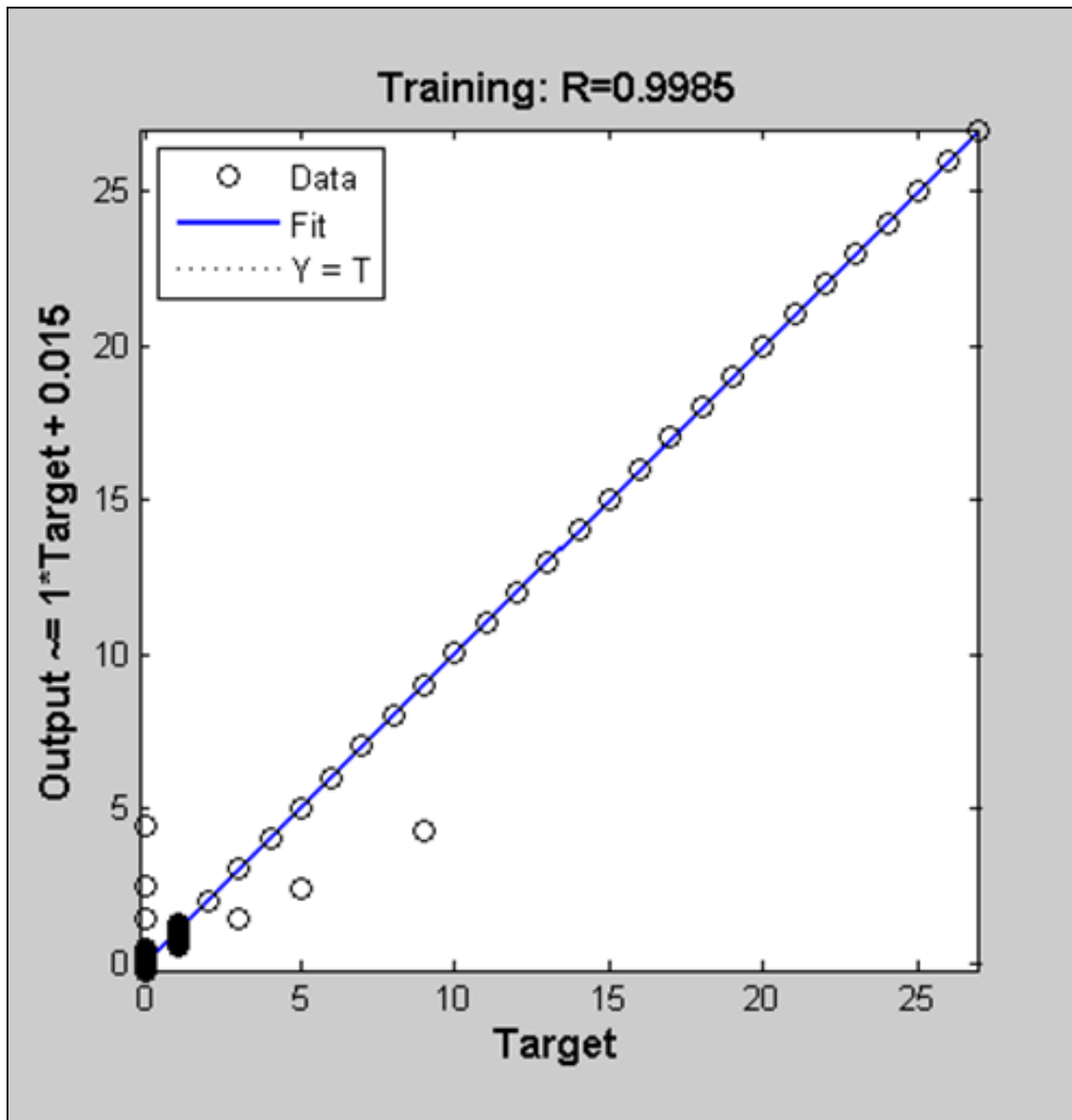


Figure 7: Typical MATLAB ANN Regression Plot

Figure 8 shows how the gradient, mu, and number of validation checks values fluctuation during the course of the network simulation. The decreasing gradient and mu

values both indicate the network was performing well. The validation checks remaining at zero also indicated that the network training was performing well. In this simulation, from Figure 3, the gradient parameter was the termination factor because it reached its lower limit.

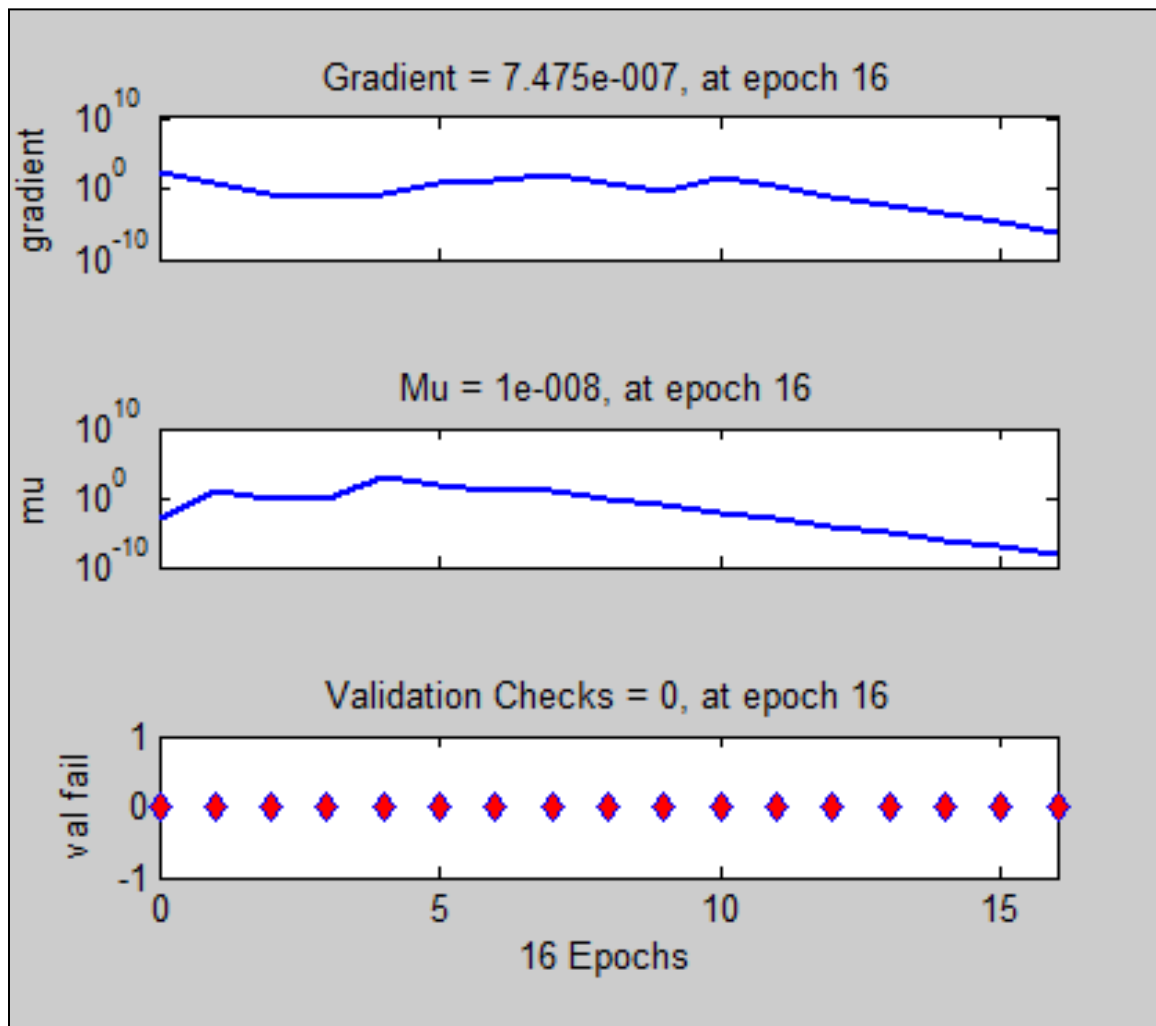


Figure 8: Typical MATLAB ANN Training States

Figure 9 shows how the gradient, mu, and number of validation checks values fluctuation during the course of the network simulation. The decreasing gradient and mu values both indicate the network was performing well. The validation checks remaining

at zero also indicated that the network training was performing well. In this simulation, from Figure 3, the gradient parameter was the termination factor because it reached its lower limit.

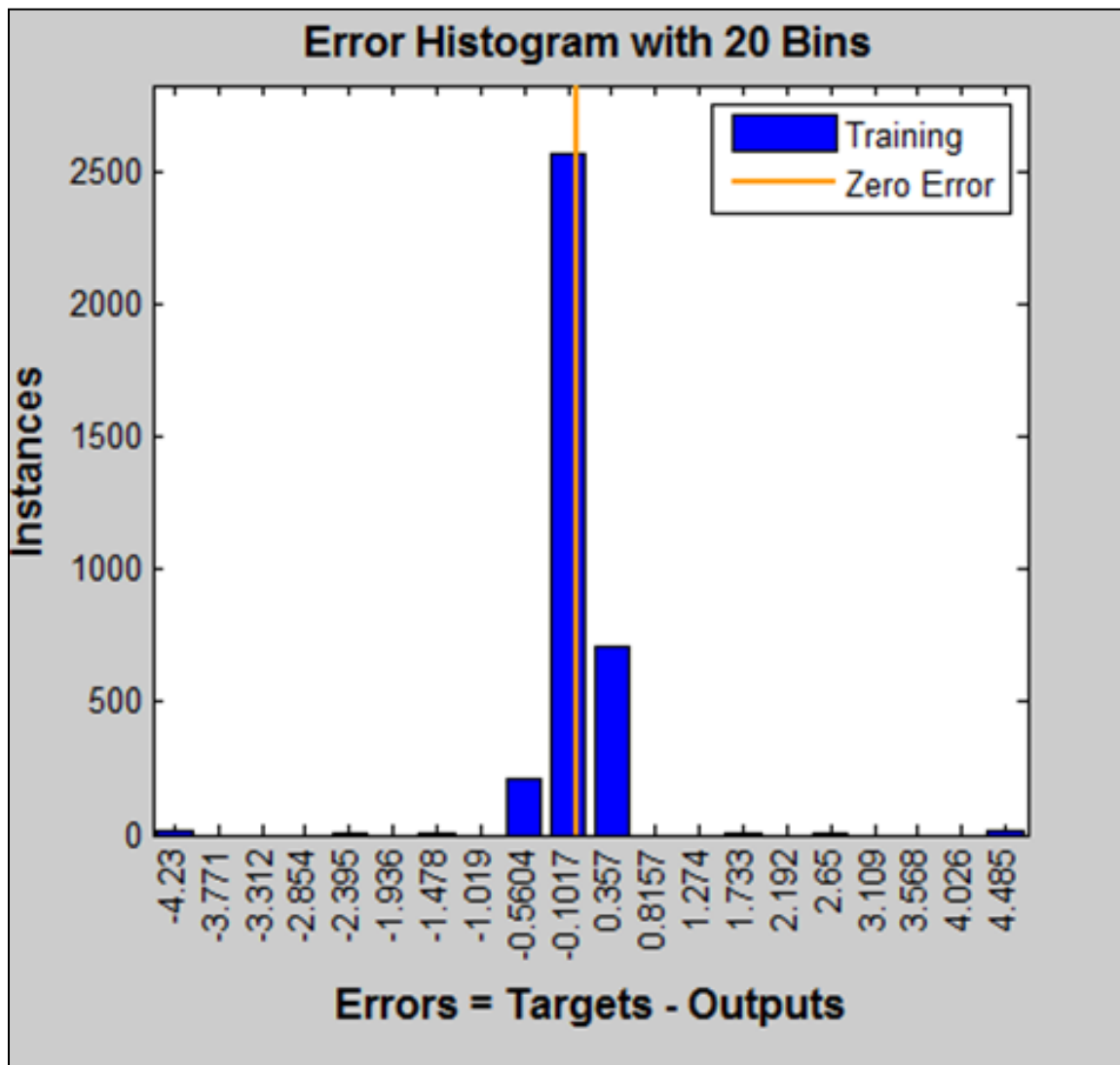


Figure 9: Typical MATLAB ANN Error Histogram Plot

After the ANN was shown to be capable of fitting chemical and disease input and output data in the model. The TVT ratios and training functions were tested in the

second phase of simulations. Using the TVT ratio as an independent variable yields a wide range of training performance parameter values. Table 5 shows the effects that different TVT ratios can have on the network performance criteria. From the data collected, the 70-15-15 TVT ratio provided the best overall performance for the training of the network. The higher average number of epochs associated with the 70-15-15 TVT ratio shows that this ratio allowed the network more opportunity to improve with more simulation iterations than the other ratios. While the 60-20-20 TV ratio had a lower average performance value, it had a higher average mu value and a higher average number of validation checks indicating that the 60-20-20 ratio did not adequately establish proper weight values and could not continue to improve performance of the network. Although the 70-15-15 TVT ratio had the highest average gradient value, it had the lowest average mu and lowest number of validation checks of the five TVT ratios. Table 5 shows that over the five network training criteria, the 70-15-15 TVT ratio provides the best overall performance with that data.

Table 5: The Effect of TVT Ratio on Network Training Statistics

Ratio		Network Training Parameters					
		Epochs	Time	Performance	Gradient	Mu	Validation Checks
50-25-25	Minimum	3	37.75	3.929	1.174	2.00E-07	1
	Maximum	8	208.05	92.510	642.100	2.00E+04	6
	Average	4.4	90.07	37.277	192.282	2.91E+03	3
60-20-20	Minimum	4	69.48	0.380	0.213	6.02E-06	4
	Maximum	12	219.66	96.450	993.201	2.40E+05	6
	Average	7.8	146.72	20.792	310.266	6.92E+04	5.2
70-15-15	Minimum	10	154.85	0.177	0.000	8.20E-07	0
	Maximum	14	256.19	90.037	2326.863	2.40E+02	0
	Average	11.6	195.68	24.871	422.535	2.00E+01	0
80-10-10	Minimum	6	92.27	16.926	0.000	6.00E-04	0
	Maximum	19	276.79	95.657	810.218	2.01E+04	0
	Average	9.8	150.64	46.503	172.010	2.12E+03	0
90-5-5	Minimum	2	32.45	0.000	0.001	4.02E-04	0
	Maximum	9	192.81	94.892	512.904	2.20E+05	6
	Average	4.4	101.67	42.661	120.958	4.69E+03	3.4

Figure 10 shows the R^2 -values for the five different TVT ratios used in network simulations. The 70-15-15 percent ratio clearly provides the best fit for the data with a high R -value of nearly one. The 80-10-10 percent ratio also performs well but has a lower R^2 -value than the 70-15-15 percent ratio. The other ratios have low R^2 -values and the plots of the ANN derived disease numbers versus actual disease numbers show the inaccuracy of those models. Individual plots for each of the TVT ratios used can be found in Appendix D.

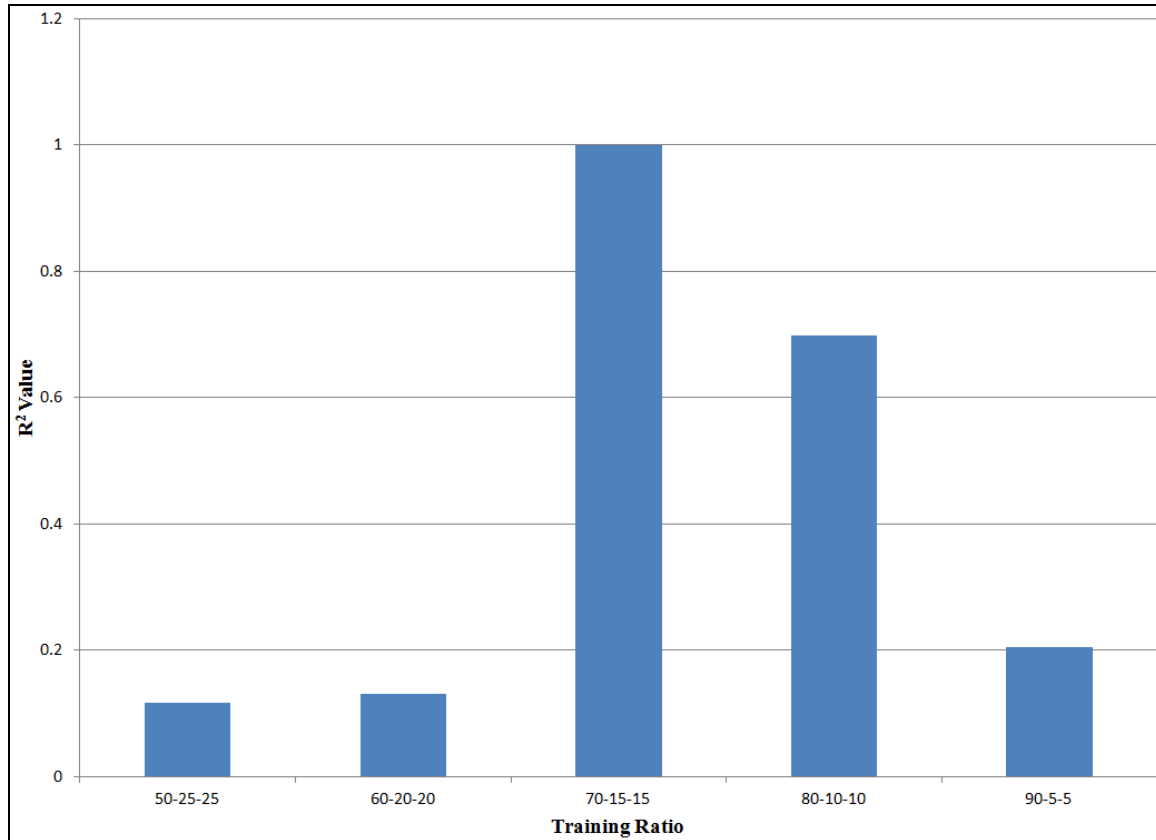


Figure 10: TVT Ratio Effect on the Coefficient of Determination

In addition to TVT ratios, altering the training functions used to train the network can also affect network performance. Table 6 shows the effects that different training functions have on the network performance criteria. When conducting simulations with the different training functions, the 70-15-15 TV ratio was used to organize the data as it was determined to provide the best performance with the network. Explanations for all of the training functions used can be found in Appendix E. From the data collected in the simulations, the trainlm provided the best overall training performance for the simulations. Because not all of the training functions used gradient, mu or validations checks as performance parameters, epochs, time and performance were the main

parameters used to compare the different functions. The maximum number of epochs was used to terminate the simulations for 10 of the 15 training functions used in the network. While the `trainlm` training function did not have the lowest performance value, its simulations were terminated due to the gradient reaching the lower limit indicating high network performance. Training function `traingdscg` had a lower performance parameter but also had a lower R-value when the actual and ANN derived disease numbers were plotted. It should also be noted that training functions with shorter network run times generally performed worse than functions with longer run times.

Table 6: The Effects of Training Functions on Network Training Statistics

Function		Performance Parameters					
		Epochs	Time	Performance	Gradient	Mu	Validation Checks
trainb	Minimum	1000	32.94	85.55	-	-	0
	Maximum	1000	47.24	102.43	-	-	0
	Average	1000	38.27	93.48	-	-	0
trainc	Minimum	1000	47.95	65.06	-	-	-
	Maximum	1000	306.49	101.05	-	-	-
	Average	1000	214.66	79.73	-	-	-
traincgb	Minimum	3	0.18	64.27	2.68	-	0
	Maximum	13	0.22	88.41	416.71	-	0
	Average	9	0.20	71.86	260.10	-	0
traincgf	Minimum	101	0.75	40.37	1.07	-	0
	Maximum	659	4.90	88.26	457.27	-	0
	Average	290	2.17	47.37	72.29	-	0
traincgp	Minimum	5	0.26	55.00	1.68	-	0
	Maximum	161	1.62	89.20	361.27	-	0
	Average	88	0.94	59.85	38.10	-	0
traingd	Minimum	1000	2.85	48.12	0.93	-	0
	Maximum	1000	2.86	100.66	229.21	-	0
	Average	1000	2.86	61.70	2.30	-	0
traingda	Minimum	1000	2.99	68.21	195.53	-	0
	Maximum	1000	3.30	146.91	408.98	-	0
	Average	1000	3.18	73.10	329.39	-	0
traingdm	Minimum	1000	3.11	48.10	0.94	-	0
	Maximum	1000	3.56	104.42	576.00	-	0
	Average	1000	3.27	84.38	2.55	-	0
traindgx	Minimum	1000	3.25	78.80	3.17	-	0
	Maximum	1000	3.27	95.30	602.00	-	0
	Average	1000	3.59	80.53	133.58	-	0
trainlm	Minimum	10	154.85	0.18	0.00	0.00	0
	Maximum	14	256.19	91.97	2143.91	400.00	0
	Average	13	219.36	27.91	394.93	33.25	0
trainoss	Minimum	1000	6.35	0.34	0.00	-	0
	Maximum	1000	10.19	99.07	550.00	-	0
	Average	1000	7.65	36.37	2.49	-	0
trainr	Minimum	1000	52.66	64.00	-	-	-
	Maximum	1000	87.99	101.85	-	-	-
	Average	1000	64.48	79.43	-	-	-
trainrp	Minimum	45	0.27	42.00	0.00	-	0
	Maximum	49	0.35	99.69	524.00	-	0
	Average	47	0.31	47.37	5.53	-	0
trains	Minimum	1000	21.25	83.90	-	-	-
	Maximum	1000	25.84	100.76	-	-	-
	Average	1000	23.35	91.88	-	-	-
trainscg	Minimum	1000	4.60	0.44	0.12	-	0
	Maximum	1000	7.86	95.46	545.60	-	0
	Average	1000	5.87	25.54	9.57	-	0

Figure 11 shows the R^2 -values for each of the training functions used in the network simulations. Several of the training functions had good R^2 -values, above 0.7, but the majority fell around or below 0.5 indicating most training functions did not have a good fit with the data. The Trainlm function had the highest R^2 -value of 0.999. Trainrp also had an R^2 -value of 0.999 but the slope of the line 0.5:1, not 1:1. Trainscg had the second highest R^2 -value of 0.9769 but the ANN derived disease number versus actual disease number plot was not as linear as Trainlm. Plots for all of the training functions can be found in the Appendix.

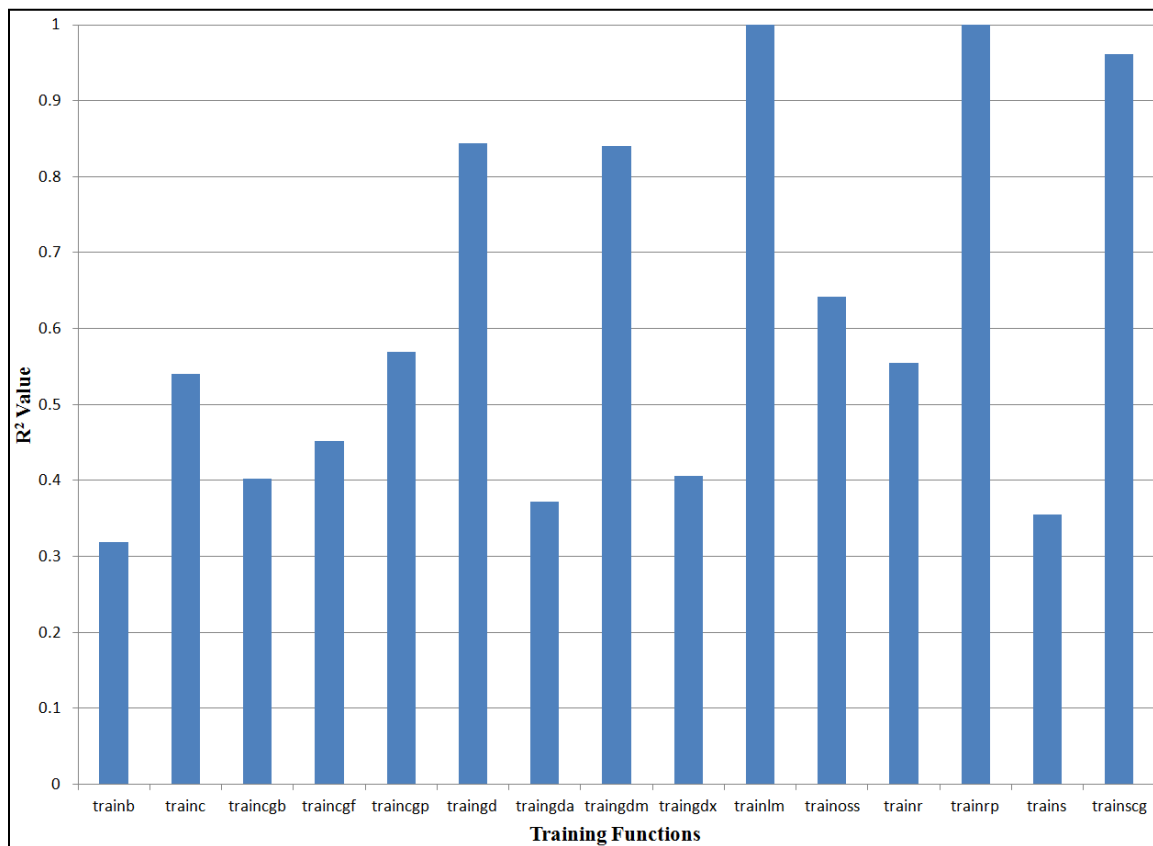


Figure 11: Training Function Effect on the Coefficient of Determination

ANN Model Performance for Curated Chemicals

Utilizing curated chemicals from the CTD allowed various network models to be created testing different TVT ratios. The actual disease number is known when using curated chemicals so the ANN derived disease numbers can easily be compared to the actual values to determine how well the network is performing.

Effect of Training Ratio on Model-Predicted Disease

Figure 12 shows typical actual disease number versus ANN derived disease number plots for each of the five TVT ratios used. For the simulations shown, default training functions and parameters were used in the network. The graph also includes a one-to-one slope line to easily compare the different simulations results to the desired values. As discussed earlier, the 70-15-15 percent TVT ratio generated the best network performance. Comparing the different TVT ratio plots, it is evident that the other ratios do not produce the same performance as the 70-15-15 percent ratio and are unable to accurately generate ANN derived disease numbers.

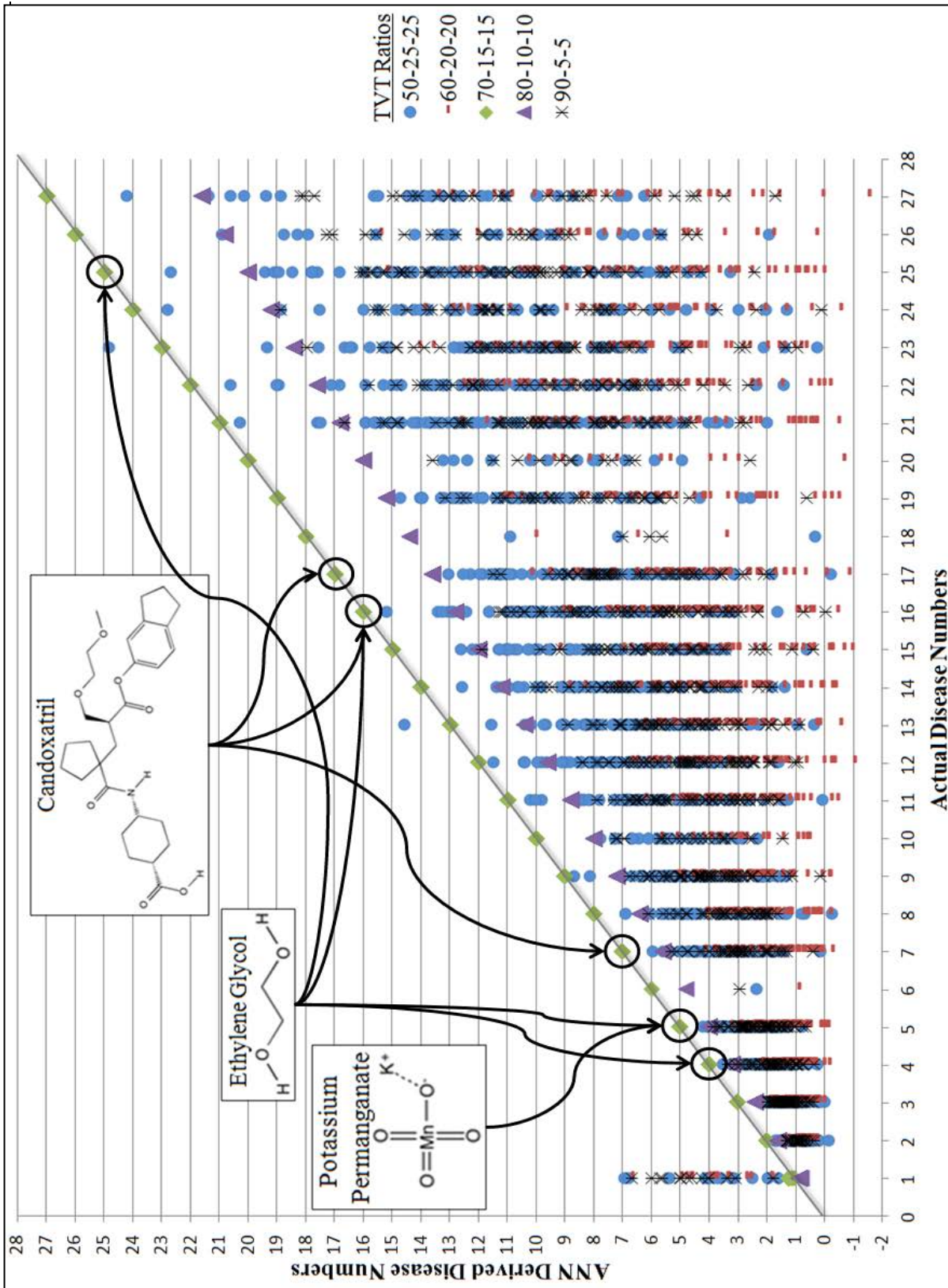


Figure 12: Effect of TVT Ratio on MATLAB ANN Derived Disease

Figure 12 also highlights three structurally different chemicals used in the network simulations showing how their ANN derived disease numbers compare to their actual disease numbers when using the 70-15-15 percent TVT ratio. Two important details about the network can be seen when examining the three highlighted chemicals. First, the network is able to take inputs from different chemicals and correctly match it to a single disease category. Potassium permanganate and ethylene glycol can both be correctly linked to disease group five and ethylene glycol and candoxatril can both be correctly linked to disease category 16. In addition to matching multiple chemicals to one disease category, the network also correctly took inputs from a single chemical and linked it to multiple diseases that it is related to. Both ethylene glycol and candoxatril are shown to correctly have associations with multiple disease categories.

Chemical Trends for Undertrained Model Simulations

Figure 13 shows the effects of using an undertrained network on the ANN derived disease number using the average of five trials of simulation data. Nearly all of the ANN derived disease numbers fall below the one to one slope line which correlates to the low R^2 -value seen in Figure 4.8. For every disease associated with each of the three chemicals, the ANN derived disease number was lower than the actual value. The network was unable to derive the same disease number for multiple chemicals that were linked to the same disease. When the network did not have enough data available to properly train with, the performance of the network suffered and produced inaccurate results.

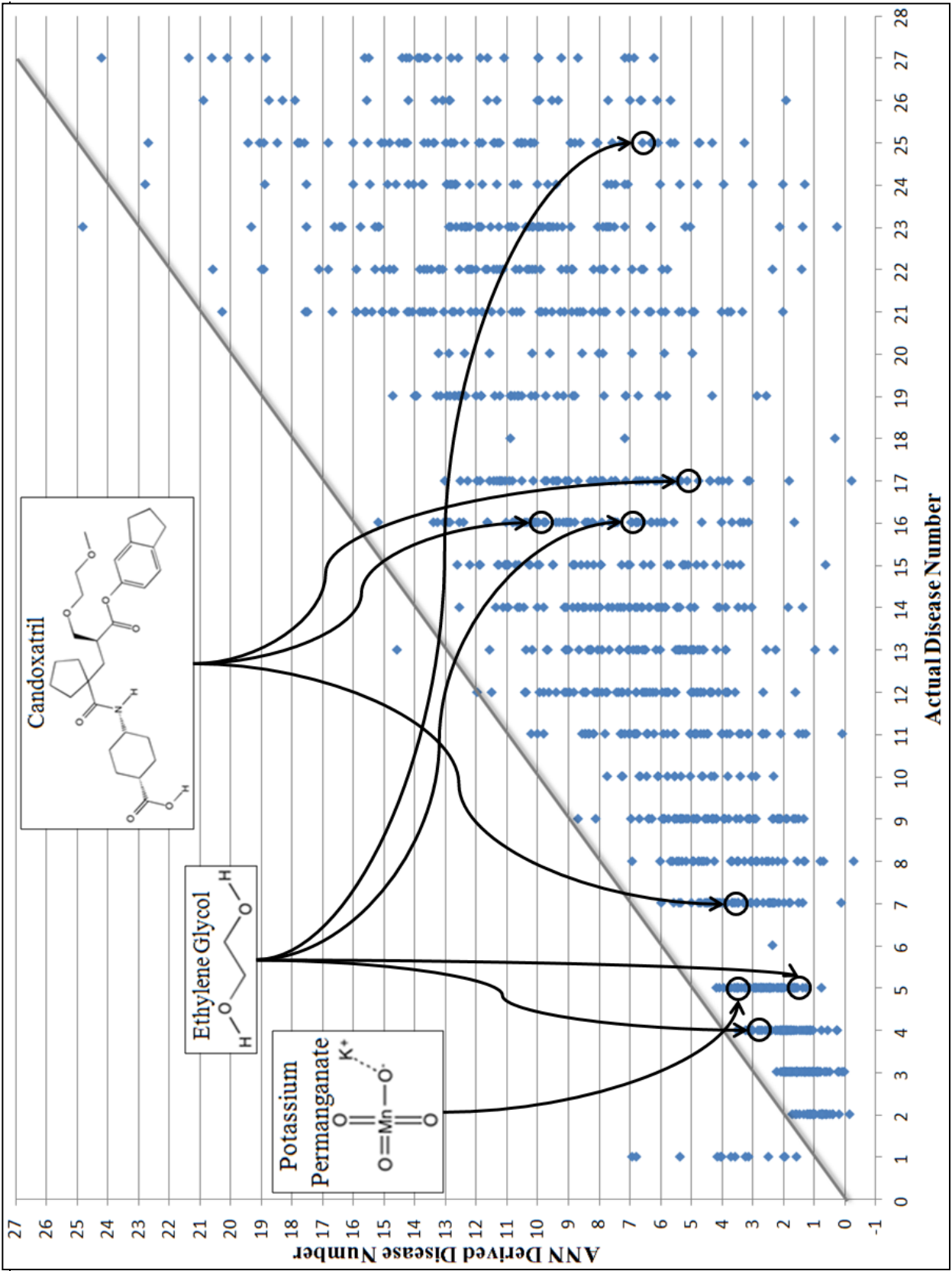


Figure 13: Effect of Undertrained TVT Ratio (50-25-25 %) on MATLAB ANN Derived Disease

Figure 13 also shows the three chemicals highlighted in Figure 12 to display how the network derives disease numbers for them when not provided with enough training data. When examining the three chemical's ANN derived disease numbers, it is evident that the undertrained network is not able to predict the correct values. When more than one chemical is linked to the same disease, the network is unable to predict the same disease for the multiple chemicals. Potassium permanganate and ethylene glycol should both be related to disease category 5 but the network predicts values near 3.5 and 1.5 respectively. Not providing enough data for the network to train with negatively affects the overall performance of the network.

Occasionally, less than 0.5% of the time, the network will generate a disease value close to the actual value, but this does not occur on a consistent basis. It is also not consistent for a certain chemical or disease. For example, in one simulation, the network generated a disease number of 14.2692 for formaldehyde. Formaldehyde is related to disease category 15 (neoplasms) so this is a difference of 4.9%. Formaldehyde is also related to 23 other disease groups and the closest network generated disease number out of the remaining 23 was 25.6% off the actual number. Out of the five simulations run with the network, this was also the only time a number within 5% of 15 was generated for formaldehyde.

Chemical Trends for Overtrained Model Simulations

Figure 14 shows the effects of using an overtrained network on the ANN derived disease number. The results shown in Figure 14 were obtained using a TVT ratio of 90-5-5 percent and the default network training settings. Similar to the results seen in Figure 13, the network was unable to generate correct disease numbers and nearly all of the generated values fell below the one to one slope line. The overtrained network saw similar failures when choosing disease number for multiple chemicals associated with one disease. Potassium permanganate and ethylene glycol should have had a disease number of five generated by the network, but instead values of 2.75 and 1.75 were generated for the two chemicals, respectively.

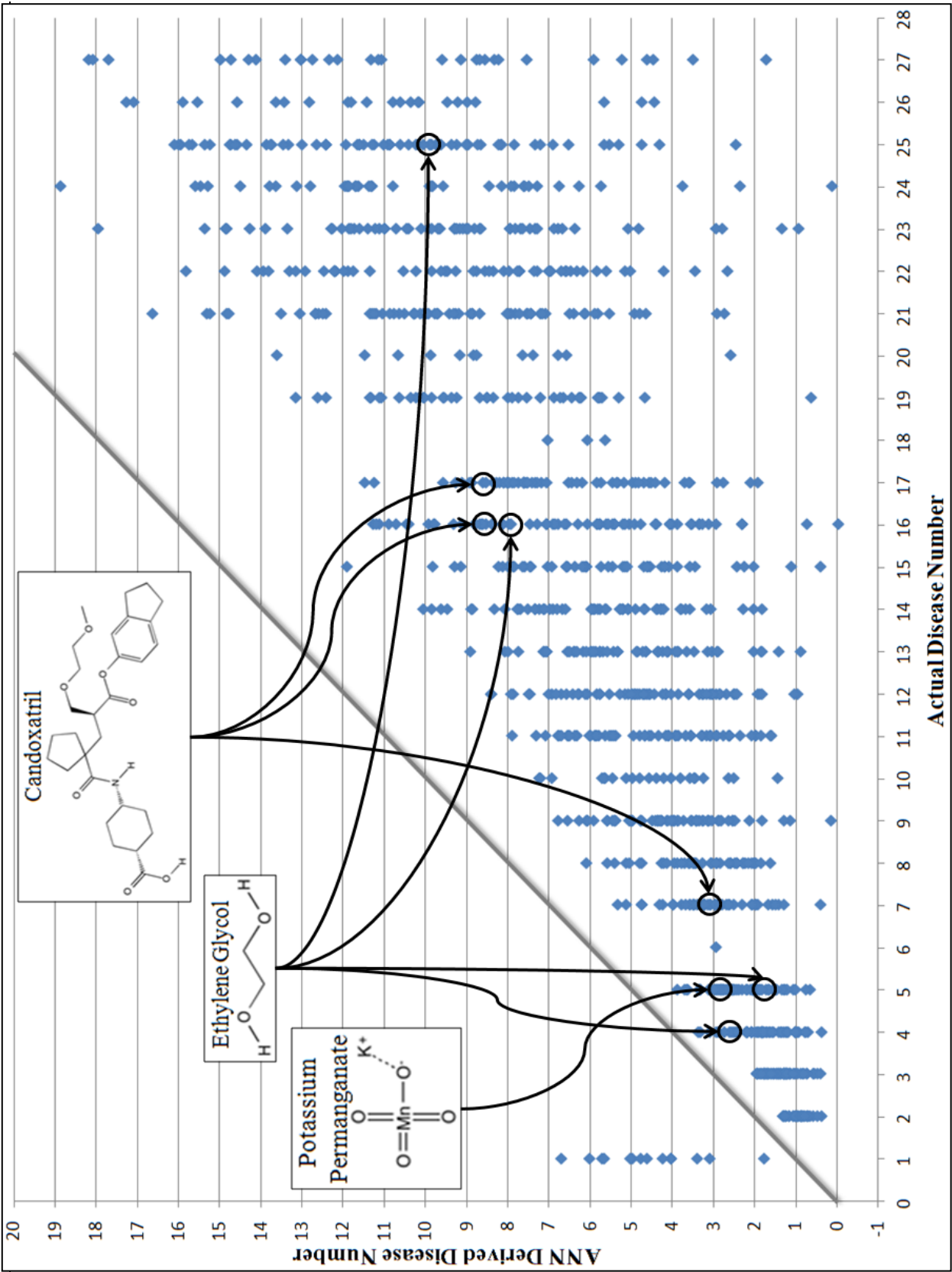


Figure 14: Effect of Overtrained TVT Ratio (90-5-5 %) on MATLAB ANN Derived Disease

ANN Model Performance for Uncurated Chemicals

Three uncurated chemicals from the CTD were selected to be tested by the ANN using a 70-15-15 TVT ratio and the trainlm training function. While these chemicals were uncurated in the database, there is research to suggest that they are related to certain diseases. Testing the network with these chemicals gives the ability to determine how well the model can predict chemical-disease associations.

Figure 15 shows the network results when the uncurated chemical cystaphos was input into the network. From the 10 inputs, the network generated seven possible disease output numbers. Of the seven predicted outputs, neoplasm was generated by the network and had literature from previous research to support this network prediction. For example, the Defense Threat Reduction Agency (DTRA) in 2006 conducted a review of previous Soviet Union research involving biological actions of neutron radiobiology. DTRA discovered several trials where animals were injected with cystaphos or cystaphos mixed with other chemicals. These trials included results that showed cystaphos was capable of protecting the intestinal system in mice from unwanted radiation damage (Defense Threat Reduction Agency, 2006). In addition to the DTRA report, Barkaia et al. conducted research in 1989 involving cystaphos as an adjuvant in cancer treatment. Using mice, guinea pigs, monkeys, they injected the animals with cystaphos combined with sodium nitrite and mexamine after irradiating them with Cs-137 gamma rays. Barkaia et al. then monitored the radiation sickness exhibited by the animals and repeated the cystaphos solution injections to see if the radiation sickness lessened. From their experiments, they found that with repeated injections, the cystaphos solution helped to

protect healthy cells in the bone marrow, spleen, and intestine of the test animals (Barkaia et al., 1989). The findings of Barkaia et al. mimic those mentioned in the DTRA report in that cystaphos appears to reduce the harmful effects of radiation and protect the intestinal system in mice. The results presented from these two studies support the network model prediction that cystaphos is linked to neoplasms. This identified potential link does not indicate that cystaphos causes cancer, but rather it is connected to it in ways that are not fully understood. The peer-reviewed literature does not contain studies that have examined the connection between cystaphos and any of the other predicted diseases.

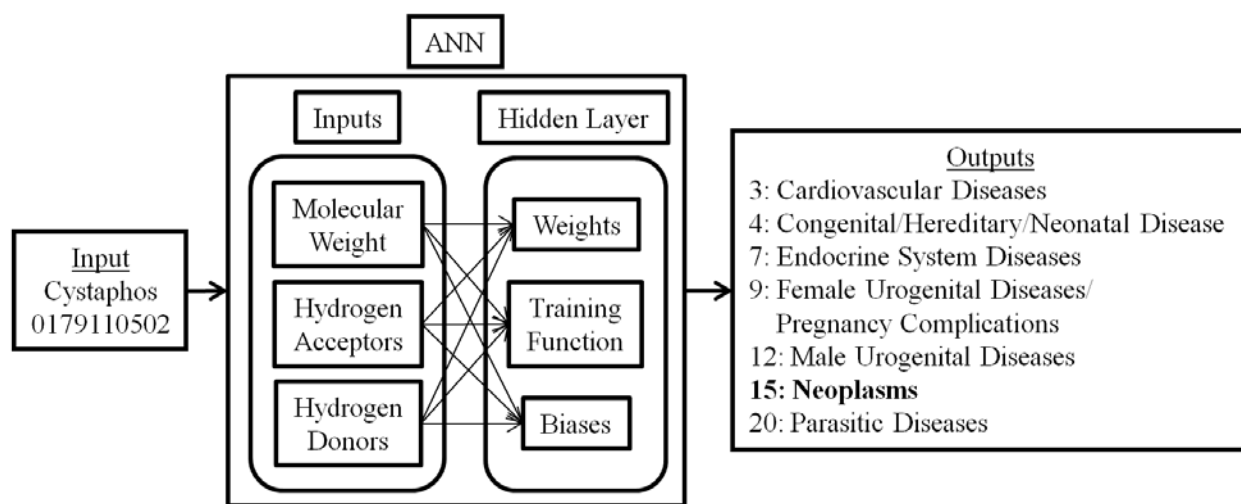


Figure 15: MATLAB ANN Derived Diseases for NCC Cystaphos

Figure 16 shows the network results when the uncured chemical 3,5-dibromo-2-(2,4-dibromophenoxy)phenol (6-HO-BDE-47) was input into the network. From the 10 inputs, the network generated nine possible disease output numbers with nervous system diseases being supported by previous research. Hendriks et al., 2010 studied the use of polybrominated diphenyl ethers (PDE) to stimulate nicotinic acetylcholine (nACh) and GABA(A) receptors on neurons in the brain. Hendriks et al. took several PDEs, including 6-HO-BDE-47, and conducted tests investigating their effects on the nACh and

GABA(A) receptors. Their findings documented that 6-HO-BDE-47 was an antagonist to the nACH receptors yet acted as an agonist to GABA(A) receptors. The results showing that 6-HO-BDE-47 can be linked to nervous systems diseases from the work of Hendriks et al. supports the same prediction generated by the ANN model. In addition to being related to nervous system diseases, there is also literature to support a possible connection to endocrine diseases. The network model did not predict endocrine diseases but an investigation conducted by Cao et al., 2010 indicates that this may be a possibility. Cao et al. took PDEs that were known to cause thyroid hormone disruption and tested to see if they bind to hormone transport proteins. Their results showed that 6-OH-PDE-47 had an affinity to binding with the thyroid hormone transport protein which could cause endocrine system problems. Although the network predictions parallel the findings of Hendriks et al. and Cao et al., the model is only able to potentially link 6-OH-PDE-47 to nervous system and endocrine system diseases. It does not indicated that 6-OH-PDE-47 directly causes these diseases. The peer-reviewed literature does not contain studies that have examined the connection between 6-HO-PDE-47 and any of the other predicted diseases.

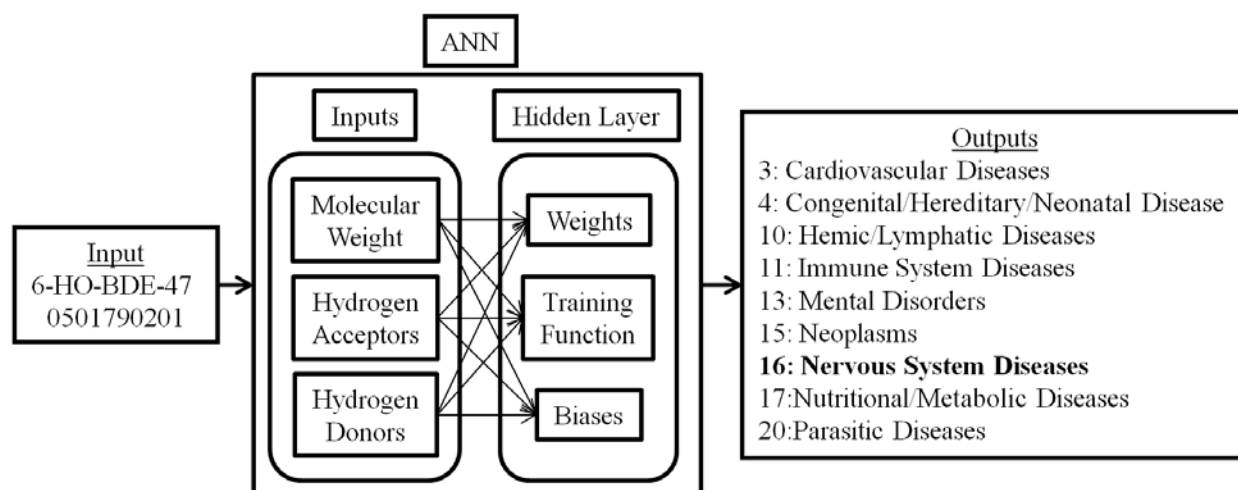


Figure 16: MATLAB ANN Derived Diseases for NCC 6-HO-BDE-47

Figure 17 shows the network results when the uncured chemical 4,4'-diiodobiphenyl (DIB) was input into the network. Seven output predictions were generated from the 10 inputs and endocrine system diseases had literature to support the network's prediction. Yomada-Okabe et al., 2005 published research findings suggesting DIB affected thyroid hormone receptors by inhibiting gene expression. From their work, Yomada-Okabe et al. concluded that DIB affects the luciferase gene by enhancing the expression of it. Mediation of the luciferase gene has been documented to act as an endocrine disruptor in animals and humans. This relationship indicates that DIB could be a potential source of endocrine disease as predicted by the model. The peer-reviewed literature does not contain studies that have examined the connection between DIB and any of the other predicted diseases.

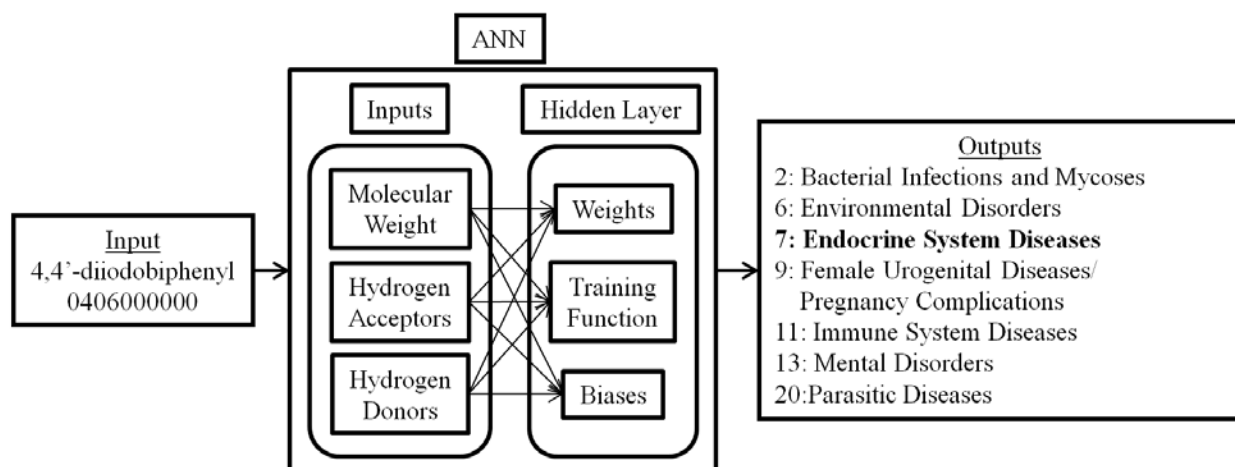


Figure 17: MATLAB ANN Derived Diseases for NCC 4,4'-diiodobiphenyl

VI. Conclusions and Recommendations

Research Conclusions

The conclusions from this research are as follows:

- 1) A new chemical classification system was created to identify chemicals with a unique number based on structural characteristics of the chemical. This new system facilitates the analysis of relationships between chemicals and diseases. While the use of molecular weight, hydrogen acceptors, and hydrogen donors proved sufficient for creating the classification system that is able to be used in a predictive ANN, the ANN model results do not prove that these three variables provide the best performance results for predicting chemical-disease associations. Other chemical characteristics may provide equal or better results.
- 2) Artificial neural networks were successfully employed to associated chemicals and diseases. Initial simulations with TVT ratios of 80-10-10 percent produced coefficients of determination equal to 0.99. The ANN derived diseases were predicted using inputs that were formatted according to the new chemical classification system.
- 3) The TVT ratio of 70-15-15 percent provided the best network performance when compared to other TVT ratios. When compared to the other ratios tested in the network, the 70-15-15 percent TVT had the lowest performance values, or lowest error values, for ratios that produced network with zero

validation checks. The lack of validation checks shows that the ANN was able to properly train the data. Additionally, the 70-15-15 percent ratio had the highest R^2 value (0.99) of the TVT ratios indicating it has a greater likelihood of accurately predicting disease output values.

- 4) The `trainlm` (Levenberg-Marquardt backpropagation) function provided the best network performance. The `trainlm` function had an R^2 -value of 0.99 and a one-to-one slope on the actual disease number versus ANN derived disease number plot. The `trainrp` and `trainscg` functions also produce good results with R^2 -values of 0.99 and 0.976 respectively; however, the `trainrp` function results did not follow a one-to-one linear slope and the `trainscg` function did not produce linear results. The high R^2 value indicates that the `trainlm` function has a greater chance of correctly predicting disease output values.
- 5) The ANN has potential to predict chemical-disease associations that are not yet curated. Cystaphos was correlated to neoplasms and two independent literature sources supported the ANN prediction. The ANN predicted that 3,5-dibromo-2-(2,4-dibromophenoxy)phenol (6-HO-BDE-47) was associated with nervous system diseases and a research documentation supported this finding. A separate literature source concluded that 6-HO-BDE-47 was also linked to endocrine diseases but the ANN failed to make that connection. 4,4'-diiodobiphenyl (DIB) was correctly matched to endocrine disease and supported by an independent research report. The ANN has demonstrated the potential to predict disease-associations for new chemicals and to guide research for existing chemicals that require toxicological testing.

Recommendations for Future Research

Future research should carry out the following activities:

- 1) Laboratory testing of chemical-disease associations that are predicted by the ANN model presented in this thesis, followed by possible refinements or modifications to the new chemical classification system. This should be carried out for both curated and uncurated chemicals. One possibility would be to analyze chemicals that were grandfathered into the Toxic Substances Control Act inventory whose chemical-disease associations are unknown.
- 2) Developing an ANN that correlates chemical-gene expression associations in order to develop a tool that provides insight into the mechanisms that cause (or prevent) disease.
- 3) Testing of additional factors in the ANN to produce a more accurate model for correlating chemicals, genes, and diseases. Utilized a loop function to iteratively step through every possible TVT ratio combination may discover a more optimal ratio to use in the network. Additionally, testing transfer functions within the ANN hidden layer or adding additional hidden layers has the potential to show increased network predictive performance as well.

Appendix A: MATLAB ANN Code

```
1  % This code is used for the thesis work conducted by Capt Brouch
2  % This code relates chemical input data to species and disease output data
3  % This code is based on the sample input-output fitting network in the MATLAB ANN Guide
4  % The script was last revised on 22 Jan 2014 by Capt Brouch
5  % The format for the input matrix is: ['Molecular Weight' 'Hydrogen Acceptors' 'Hydrogen Donors']
6  % The format for the output matrix is: ['Species' 'Dummy Variable' 'Disease']
7  % The format for the input1 matrix is: ['Molecular Weight' 'Hydrogen Acceptors' 'Hydrogen Donors']
8  % The input1 matrix contains uncurated chemical data used to predict disease outputs
9
10- tstart = clock;
11
12- A = zeros(1173,3);
13- B = zeros(1173,3);
14- C = zeros(1173,3);
15
16- A(:,1) = xlsread('J:\Brouch\Brouch Thesis\Excel Files\Brouch Thesis ANN Data.xls','Input-Output Tables','b2:b1174');
17- A(:,2) = xlsread('J:\Brouch\Brouch Thesis\Excel Files\Brouch Thesis ANN Data.xls','Input-Output Tables','c2:c1174');
18- A(:,3) = xlsread('J:\Brouch\Brouch Thesis\Excel Files\Brouch Thesis ANN Data.xls','Input-Output Tables','d2:d1174');
19- B(:,1) = xlsread('J:\Brouch\Brouch Thesis\Excel Files\Brouch Thesis ANN Data.xls','Input-Output Tables','e2:e1174');
20- B(:,3) = xlsread('J:\Brouch\Brouch Thesis\Excel Files\Brouch Thesis ANN Data.xls','Input-Output Tables','g2:g1174');
21- C(:,1) = xlsread('J:\Brouch\Brouch Thesis\Excel Files\Brouch Thesis ANN Data.xls','Uncurated','c3:c1175');
22- C(:,2) = xlsread('J:\Brouch\Brouch Thesis\Excel Files\Brouch Thesis ANN Data.xls','Uncurated','d3:d1175');
23- C(:,3) = xlsread('J:\Brouch\Brouch Thesis\Excel Files\Brouch Thesis ANN Data.xls','Uncurated','e3:e1175');
24
25- inputs = A;
26- targets = B;
27- inputs1 = C;
28- % the targets matrix is the same as the output matrix
29
30 % preallocate the plotting matrix (PM)
31- PM = zeros(1173,5);
32
33 % the variable is interest is vv - this is the column number in the target matrix
34- vv = 3;
35
36- countt = 1;
37
38- for count = 1:1:5
39 % Create a Fitting Network
40- hiddenLayerSize = 1;
41- net = fitnet(hiddenLayerSize);
42 % Set up Division of Data for Training, Validation, Testing
43- net.divideParam.trainRatio = 70/100;
44- net.divideParam.valRatio = 15/100;
45- net.divideParam.testRatio = 15/100;
```

```

46 % Train the Network
47- [net,tr] = trainlm(net,inputs,targets);
48 % Test the Network
49- outputs = net(inputs);
50- errors = gsubtract(outputs,targets);
51- performance = perform(net,targets,outputs);
52 % View the Network
53 % view(net)
54 % plotperf(tr)
55
56- Outputs = net(inputs);
57 % trOut = Outputs(tr.trainInd);
58 % vOut = Outputs(tr.valInd);
59 % tsOut = Outputs(tr.testInd);
60 % trTarg = targets(tr.trainInd);
61 % vTarg = targets(tr.valInd);
62 % tsTarg = targets(tr.testInd);
63 % figure (98)
64 % plotregression(trTarg,trOut,'Train',vTarg,vOut,'Validation',tsTarg,tsOut,'Testing');
65
66- PM(:,countt) = Outputs(:,vv);
67- countt = countt +1;
68- end
69
70- figure(1)
71 %plot(targets(:,vv),PM(:,1),'ro')
72 %plot(targets(:,vv),PM(:,1),'ro',targets(:,vv),PM(:,2),'go',targets(:,vv),PM(:,3),'ko', targets(:,vv),PM(:,4),'m--',
    targets(:,vv),PM(:,5),'k-.')
73- title('The relationship between actual and predicted diseases, version 1')
74- xlabel('Actual Disease Number')
75- ylabel('ANN-Derived Disease Number')
76
77- tstop = clock;
78- runtime = etime (tstop,tstart)/60;
79- disp('length of run in minutes:')
80- disp(runtime)

```

This MATLAB code was used to set up and run the various ANN simulations discussed in the paper. The first eight lines of code explain what the code provide some background information about what the code is being used for and how the input and output data is organized. Line 10 begins a clock to track how long the ANN simulations take to complete. Lines 12-14 create matrices for the input and output data obtain from Microsoft Excel spreadsheets. Lines 15-22 dictate where MATLAB fill find the data needed for the input and output matrices. Lines 16-18 and 19-20 state the data needed for the input and output matrices, respectively. Lines 21-23 identify the uncured chemical data used to generate the ANN derived disease predictions. Lines 25-27 define the input, output, and output1 matrices created in lines 12-14. Line 31 creates the plotting matrix where the ANN derive diseases values will be saved. In this code, the matrix is created with five columns so the five simulations can be run back to back. Line 34 designates the variable of interest to be saved in the plotting matrix. For this code, the variable of the interest is the third column in the target matrix: disease. Lines 38-68 control and run the actual ANN. Line 38 establishes how many simulations the network will run. For this code, 5 simulations are run to match the number of columns in the plotting matrix. Line 40 controls the number of hidden layers the network used. Lines 43-45 control how the network splits up the data according to the TVT ratio being used in the simulation. Line 47 designates the training function being used in the network. Line 56 establishes what input data the network will use to generate the ANN derive disease outputs. Lines 66-68 tell the network to stop running simulations when it reaches the predetermined maximum number established in line 38. Lines 70-75 generate the actual disease number versus ANN derived disease number plot once the network simulations are complete. Lines 77-

80 stop the clock that was started in line 10 and records the total time it took the network to run all of the simulations.

For the simulations testing the different TVT ratios and training functions, only lines 16-20 were used to obtain the input and output data. The input data for uncurated chemicals was not used in these simulations. When adjusting the TVT ratios and training functions, lines 43-45 and 47 were the only lines of code that required editing.

When the uncurated chemical data was used to generate disease value predictions, the code was first run using the original input and output to establish the network using the curate data. Then the command “`outputs=net(inputs1)`” was entered into the MATLAB command window. This told the network to use the new inputs containing the uncurated to derive disease value predicts.

Appendix B: Input and Output Matrices

Chemical	Molecular Weight	Hydrogen Acceptors	Hydrogen Donors	Species Number	Dummy Variable Column	Disease Number
Acetone	58.08	1	1	1	0	9
	58.08	1	1	1	0	12
	58.08	1	1	1	0	15
	58.08	1	1	1	0	16
	58.08	1	1	1	0	17
	58.08	1	1	1	0	25
Aciclovir	225.21	8	4	1	0	2
	225.21	8	4	1	0	3
	225.21	8	4	1	0	4
	225.21	8	4	1	0	5
	225.21	8	4	1	0	8
	225.21	8	4	1	0	9
	225.21	8	4	1	0	11
	225.21	8	4	1	0	12
	225.21	8	4	1	0	13
	225.21	8	4	1	0	15
	225.21	8	4	1	0	16
	225.21	8	4	1	0	17
	225.21	8	4	1	0	19
	225.21	8	4	1	0	20
	225.21	8	4	1	0	21
	225.21	8	4	1	0	23
	225.21	8	4	1	0	24
	225.21	8	4	1	0	25
	225.21	8	4	1	0	26
Alprazolam	308.77	4	0	1	0	3
	308.77	4	0	1	0	4
	308.77	4	0	1	0	5
	308.77	4	0	1	0	7
	308.77	4	0	1	0	8
	308.77	4	0	1	0	9
	308.77	4	0	1	0	11
	308.77	4	0	1	0	12
	308.77	4	0	1	0	13
	308.77	4	0	1	0	16
	308.77	4	0	1	0	21
	308.77	4	0	1	0	23
	308.77	4	0	1	0	25
Ammonium Sulfate	132.14	4	2	1	0	3
	132.14	4	2	1	0	17
	132.14	4	2	1	0	22
Aspirin	180.16	4	1	1	0	2

	180.16	4	1	1	0	3
	180.16	4	1	1	0	4
	180.16	4	1	1	0	5
	180.16	4	1	1	0	6
	180.16	4	1	1	0	7
	180.16	4	1	1	0	8
	180.16	4	1	1	0	9
	180.16	4	1	1	0	10
	180.16	4	1	1	0	11
	180.16	4	1	1	0	12
	180.16	4	1	1	0	13
	180.16	4	1	1	0	14
	180.16	4	1	1	0	15
	180.16	4	1	1	0	16
	180.16	4	1	1	0	17
	180.16	4	1	1	0	19
	180.16	4	1	1	0	21
	180.16	4	1	1	0	22
	180.16	4	1	1	0	23
	180.16	4	1	1	0	24
	180.16	4	1	1	0	25
	180.16	4	1	1	0	26
	180.16	4	1	1	0	27
Atenolol	266.34	5	4	1	0	3
	266.34	5	4	1	0	4
	266.34	5	4	1	0	5
	266.34	5	4	1	0	7
	266.34	5	4	1	0	9
	266.34	5	4	1	0	11
	266.34	5	4	1	0	12
	266.34	5	4	1	0	13
	266.34	5	4	1	0	14
	266.34	5	4	1	0	15
	266.34	5	4	1	0	16
	266.34	5	4	1	0	17
	266.34	5	4	1	0	21
	266.34	5	4	1	0	23
	266.34	5	4	1	0	24
	266.34	5	4	1	0	25

Azithromycin	749.00	14	5	1	0	2
	749.00	14	5	1	0	3
	749.00	14	5	1	0	5
	749.00	14	5	1	0	7
	749.00	14	5	1	0	8
	749.00	14	5	1	0	9
	749.00	14	5	1	0	10
	749.00	14	5	1	0	11
	749.00	14	5	1	0	12
	749.00	14	5	1	0	13
	749.00	14	5	1	0	14
	749.00	14	5	1	0	15
	749.00	14	5	1	0	16
	749.00	14	5	1	0	17
	749.00	14	5	1	0	19
	749.00	14	5	1	0	21
	749.00	14	5	1	0	22
	749.00	14	5	1	0	23
	749.00	14	5	1	0	24
	749.00	14	5	1	0	24
	749.00	14	5	1	0	26
Benzene	78.12	0	0	5	0	1
	78.12	0	0	1	0	3
	78.12	0	0	1	0	4
	78.12	0	0	1	0	5
	78.12	0	0	1	0	7
	78.12	0	0	1	0	9
	78.12	0	0	1	0	10
	78.12	0	0	1	0	11
	78.12	0	0	1	0	14
	78.12	0	0	1	0	15
	78.12	0	0	1	0	16
	78.12	0	0	1	0	17
	78.12	0	0	1	0	18
	78.12	0	0	1	0	21
	78.12	0	0	1	0	22
	78.12	0	0	1	0	23
	78.12	0	0	1	0	25
	78.12	0	0	1	0	27

Benzyl-penicillin	334.40	6	2	9	0	1
	334.40	6	2	1	0	2
	334.40	6	2	1	0	3
	334.40	6	2	1	0	4
	334.40	6	2	1	0	5
	334.40	6	2	1	0	7
	334.40	6	2	1	0	8
	334.40	6	2	1	0	9
	334.40	6	2	1	0	10
	334.40	6	2	1	0	11
	334.40	6	2	1	0	12
	334.40	6	2	1	0	14
	334.40	6	2	1	0	16
	334.40	6	2	1	0	17
	334.40	6	2	1	0	19
	334.40	6	2	1	0	20
	334.40	6	2	1	0	21
	334.40	6	2	1	0	22
	334.40	6	2	1	0	23
	334.40	6	2	1	0	24
	334.40	6	2	1	0	25
	334.40	6	2	1	0	26
Caffeine	194.19	6	0	9	0	1
	194.19	6	0	1	0	2
	194.19	6	0	1	0	3
	194.19	6	0	1	0	4
	194.19	6	0	1	0	5
	194.19	6	0	1	0	7
	194.19	6	0	1	0	8
	194.19	6	0	1	0	9
	194.19	6	0	1	0	10
	194.19	6	0	1	0	11
	194.19	6	0	1	0	12
	194.19	6	0	1	0	13
	194.19	6	0	1	0	14
	194.19	6	0	1	0	15
	194.19	6	0	1	0	16
	194.19	6	0	1	0	17
	194.19	6	0	1	0	19
	194.19	6	0	1	0	21
	194.19	6	0	1	0	22
	194.19	6	0	1	0	23
	194.19	6	0	1	0	25
	194.19	6	0	1	0	27
Cadoxatril	515.65	8	2	1	0	3
	515.65	8	2	1	0	7
	515.65	8	2	1	0	16
	515.65	8	2	1	0	17

Carbamazepine	236.28	3	2	9	0	1
	236.28	3	2	1	0	3
	236.28	3	2	1	0	4
	236.28	3	2	1	0	5
	236.28	3	2	1	0	7
	236.28	3	2	1	0	8
	236.28	3	2	1	0	9
	236.28	3	2	1	0	10
	236.28	3	2	1	0	11
	236.28	3	2	1	0	12
	236.28	3	2	1	0	13
	236.28	3	2	1	0	14
	236.28	3	2	1	0	15
	236.28	3	2	1	0	16
	236.28	3	2	1	0	17
	236.28	3	2	1	0	19
	236.28	3	2	1	0	21
	236.28	3	2	1	0	22
	236.28	3	2	1	0	23
	236.28	3	2	1	0	24
	236.28	3	2	1	0	25
	236.28	3	2	1	0	26
Caustic Soda	40.00	1	1	9	0	1
	40.00	1	1	1	0	5
	40.00	1	1	1	0	19
	40.00	1	1	1	0	22
	40.00	1	1	1	0	27
Chloramphenicol	323.14	7	3	1	0	2
	323.14	7	3	1	0	3
	323.14	7	3	1	0	5
	323.14	7	3	1	0	8
	323.14	7	3	1	0	9
	323.14	7	3	1	0	10
	323.14	7	3	1	0	11
	323.14	7	3	1	0	12
	323.14	7	3	1	0	14
	323.14	7	3	1	0	15
	323.14	7	3	1	0	16
	323.14	7	3	1	0	17
	323.14	7	3	1	0	19
	323.14	7	3	1	0	21
	323.14	7	3	1	0	22
	323.14	7	3	1	0	23
	323.14	7	3	1	0	25

Cimetidine	252.34	6	3	1	0	2
	252.34	6	3	1	0	3
	252.34	6	3	1	0	4
	252.34	6	3	1	0	5
	252.34	6	3	1	0	7
	252.34	6	3	1	0	8
	252.34	6	3	1	0	9
	252.34	6	3	1	0	11
	252.34	6	3	1	0	12
	252.34	6	3	1	0	13
	252.34	6	3	1	0	14
	252.34	6	3	1	0	15
	252.34	6	3	1	0	16
	252.34	6	3	1	0	17
	252.34	6	3	1	0	19
	252.34	6	3	1	0	21
	252.34	6	3	1	0	22
	252.34	6	3	1	0	23
	252.34	6	3	1	0	24
	252.34	6	3	1	0	25
	252.34	6	3	1	0	26
	252.34	6	3	1	0	27
Clonidine	230.10	3	2	1	0	3
	230.10	3	2	1	0	4
	230.10	3	2	1	0	5
	230.10	3	2	1	0	7
	230.10	3	2	1	0	8
	230.10	3	2	1	0	9
	230.10	3	2	1	0	11
	230.10	3	2	1	0	12
	230.10	3	2	1	0	13
	230.10	3	2	1	0	14
	230.10	3	2	1	0	16
	230.10	3	2	1	0	17
	230.10	3	2	1	0	19
	230.10	3	2	1	0	21
	230.10	3	2	1	0	22
	230.10	3	2	1	0	23
	230.10	3	2	1	0	24
	230.10	3	2	1	0	25

Copper Sulfate	159.61	4	0	1	0	4
	159.61	4	0	1	0	5
	159.61	4	0	1	0	9
	159.61	4	0	1	0	12
	159.61	4	0	1	0	13
	159.61	4	0	1	0	14
	159.61	4	0	1	0	16
	159.61	4	0	1	0	17
	159.61	4	0	1	0	19
	159.61	4	0	1	0	21
	159.61	4	0	1	0	22
	159.61	4	0	1	0	25
Cyclosporine	1202.64	23	5	9	0	1
	1202.64	23	5	1	0	2
	1202.64	23	5	1	0	3
	1202.64	23	5	1	0	4
	1202.64	23	5	1	0	5
	1202.64	23	5	1	0	7
	1202.64	23	5	1	0	8
	1202.64	23	5	1	0	9
	1202.64	23	5	1	0	10
	1202.64	23	5	1	0	11
	1202.64	23	5	1	0	12
	1202.64	23	5	1	0	13
	1202.64	23	5	1	0	14
	1202.64	23	5	1	0	15
	1202.64	23	5	1	0	16
	1202.64	23	5	1	0	17
	1202.64	23	5	1	0	19
	1202.64	23	5	1	0	20
	1202.64	23	5	1	0	21
	1202.64	23	5	1	0	22
	1202.64	23	5	1	0	23
	1202.64	23	5	1	0	23
	1202.64	23	5	1	0	25
	1202.64	23	5	1	0	26

Desipramine	266.39	2	1	1	0	2
	266.39	2	1	1	0	3
	266.39	2	1	1	0	4
	266.39	2	1	1	0	5
	266.39	2	1	1	0	7
	266.39	2	1	1	0	8
	266.39	2	1	1	0	9
	266.39	2	1	1	0	11
	266.39	2	1	1	0	12
	266.39	2	1	1	0	13
	266.39	2	1	1	0	15
	266.39	2	1	1	0	16
	266.39	2	1	1	0	17
	266.39	2	1	1	0	21
	266.39	2	1	1	0	22
	266.39	2	1	1	0	23
	266.39	2	1	1	0	25
Dexamethasone	392.47	5	3	1	0	2
	392.47	5	3	1	0	3
	392.47	5	3	1	0	4
	392.47	5	3	1	0	5
	392.47	5	3	1	0	7
	392.47	5	3	1	0	8
	392.47	5	3	1	0	9
	392.47	5	3	1	0	10
	392.47	5	3	1	0	11
	392.47	5	3	1	0	12
	392.47	5	3	1	0	13
	392.47	5	3	1	0	14
	392.47	5	3	1	0	15
	392.47	5	3	1	0	16
	392.47	5	3	1	0	17
	392.47	5	3	1	0	19
	392.47	5	3	1	0	20
	392.47	5	3	1	0	21
	392.47	5	3	1	0	22
	392.47	5	3	1	0	23
	392.47	5	3	1	0	24
	392.47	5	3	1	0	25
	392.47	5	3	1	0	26
	392.47	5	3	1	0	27

Diazepam	284.75	3	0	1	0	2
	284.75	3	0	1	0	2
	284.75	3	0	1	0	4
	284.75	3	0	1	0	5
	284.75	3	0	1	0	8
	284.75	3	0	1	0	9
	284.75	3	0	1	0	11
	284.75	3	0	1	0	12
	284.75	3	0	1	0	13
	284.75	3	0	1	0	14
	284.75	3	0	1	0	15
	284.75	3	0	1	0	16
	284.75	3	0	1	0	17
	284.75	3	0	1	0	19
	284.75	3	0	1	0	21
	284.75	3	0	1	0	22
	284.75	3	0	1	0	23
Diclofenac	284.75	3	0	1	0	24
	284.75	3	0	1	0	25
	284.75	3	0	1	0	27
	296.15	3	2	1	0	2
	296.15	3	2	1	0	3
	296.15	3	2	1	0	4
	296.15	3	2	1	0	5
	296.15	3	2	1	0	7
	296.15	3	2	1	0	8
	296.15	3	2	1	0	9
	296.15	3	2	1	0	11
	296.15	3	2	1	0	12
	296.15	3	2	1	0	13
	296.15	3	2	1	0	14
	296.15	3	2	1	0	15
	296.15	3	2	1	0	16
	296.15	3	2	1	0	17
	296.15	3	2	1	0	19
	296.15	3	2	1	0	21
	296.15	3	2	1	0	22
	296.15	3	2	1	0	23
	296.15	3	2	1	0	25
	296.15	3	2	1	0	26
	296.15	3	2	1	0	27

Diltiazem-HCl	414.53	6	0	1	0	3
	414.53	6	0	1	0	4
	414.53	6	0	1	0	4
	414.53	6	0	1	0	7
	414.53	6	0	1	0	8
	414.53	6	0	1	0	9
	414.53	6	0	1	0	10
	414.53	6	0	1	0	11
	414.53	6	0	1	0	12
	414.53	6	0	1	0	13
	414.53	6	0	1	0	14
	414.53	6	0	1	0	15
	414.53	6	0	1	0	16
	414.53	6	0	1	0	17
	414.53	6	0	1	0	21
	414.53	6	0	1	0	23
	414.53	6	0	1	0	25
Doxorubicin	543.53	12	7	9	0	1
	543.53	12	7	1	0	2
	543.53	12	7	1	0	3
	543.53	12	7	1	0	4
	543.53	12	7	1	0	5
	543.53	12	7	1	0	7
	543.53	12	7	1	0	8
	543.53	12	7	1	0	9
	543.53	12	7	1	0	10
	543.53	12	7	1	0	11
	543.53	12	7	1	0	12
	543.53	12	7	1	0	13
	543.53	12	7	1	0	14
	543.53	12	7	1	0	15
	543.53	12	7	1	0	16
	543.53	12	7	1	0	17
	543.53	12	7	1	0	19
	543.53	12	7	1	0	21
	543.53	12	7	1	0	22
	543.53	12	7	1	0	23
	543.53	12	7	1	0	24
	543.53	12	7	1	0	25
	543.53	12	7	1	0	26
	543.53	12	7	1	0	27
Enalaprilat	376.46	7	2	1	0	3
	376.46	7	2	1	0	5
	376.46	7	2	1	0	21

Erythromycin	733.95	14	5	9	0	1
	733.95	14	5	1	0	2
	733.95	14	5	1	0	3
	733.95	14	5	1	0	4
	733.95	14	5	1	0	5
	733.95	14	5	1	0	7
	733.95	14	5	1	0	8
	733.95	14	5	1	0	9
	733.95	14	5	1	0	10
	733.95	14	5	1	0	11
	733.95	14	5	1	0	12
	733.95	14	5	1	0	13
	733.95	14	5	1	0	14
	733.95	14	5	1	0	15
	733.95	14	5	1	0	16
	733.95	14	5	1	0	17
	733.95	14	5	1	0	19
	733.95	14	5	1	0	20
	733.95	14	5	1	0	21
	733.95	14	5	1	0	22
	733.95	14	5	1	0	23
	733.95	14	5	1	0	24
	733.95	14	5	1	0	25
	733.95	14	5	1	0	26
Ethylene Glycol	62.07	2	2	1	0	4
	62.07	2	2	1	0	5
	62.07	2	2	1	0	9
	62.07	2	2	1	0	12
	62.07	2	2	1	0	14
	62.07	2	2	1	0	16
	62.07	2	2	1	0	25

Famotidine	337.45	9	8	1	0	3
	337.45	9	8	1	0	4
	337.45	9	8	1	0	5
	337.45	9	8	1	0	7
	337.45	9	8	1	0	9
	337.45	9	8	1	0	11
	337.45	9	8	1	0	12
	337.45	9	8	1	0	13
	337.45	9	8	1	0	14
	337.45	9	8	1	0	15
	337.45	9	8	1	0	16
	337.45	9	8	1	0	17
	337.45	9	8	1	0	21
	337.45	9	8	1	0	22
	337.45	9	8	1	0	23
	337.45	9	8	1	0	25
	337.45	9	8	1	0	26
	337.45	9	8	1	0	27
Felodipine	384.26	5	1	1	0	3
	384.26	5	1	1	0	9
	384.26	5	1	1	0	12
	384.26	5	1	1	0	16
	384.26	5	1	1	0	21
	384.26	5	1	1	0	23
	384.26	5	1	1	0	24
Ferric Chloride	162.20	0	0	1	0	3
	162.20	0	0	1	0	5
	162.20	0	0	1	0	9
	162.20	0	0	1	0	11
	162.20	0	0	1	0	12
	162.20	0	0	1	0	16
	162.20	0	0	1	0	21
	162.20	0	0	1	0	25
	162.20	0	0	1	0	27

Fluorouracil	130.08	4	2	1	0	2
	130.08	4	2	1	0	3
	130.08	4	2	1	0	4
	130.08	4	2	1	0	5
	130.08	4	2	1	0	7
	130.08	4	2	1	0	8
	130.08	4	2	1	0	9
	130.08	4	2	1	0	10
	130.08	4	2	1	0	11
	130.08	4	2	1	0	12
	130.08	4	2	1	0	13
	130.08	4	2	1	0	14
	130.08	4	2	1	0	15
	130.08	4	2	1	0	16
	130.08	4	2	1	0	17
	130.08	4	2	1	0	19
	130.08	4	2	1	0	21
	130.08	4	2	1	0	22
	130.08	4	2	1	0	23
	130.08	4	2	1	0	24
	130.08	4	2	1	0	25
	130.08	4	2	1	0	26
	130.08	4	2	1	0	27
Flurbiprofen	244.27	2	1	1	0	3
	244.27	2	1	1	0	4
	244.27	2	1	1	0	5
	244.27	2	1	1	0	8
	244.27	2	1	1	0	9
	244.27	2	1	1	0	11
	244.27	2	1	1	0	12
	244.27	2	1	1	0	13
	244.27	2	1	1	0	14
	244.27	2	1	1	0	15
	244.27	2	1	1	0	16
	244.27	2	1	1	0	17
	244.27	2	1	1	0	19
	244.27	2	1	1	0	21
	244.27	2	1	1	0	22
	244.27	2	1	1	0	23
	244.27	2	1	1	0	24
	244.27	2	1	1	0	25

Formaldehyde	30.03	1	0	3	0	1
	30.03	1	0	5	0	1
	30.03	1	0	1	0	2
	30.03	1	0	1	0	3
	30.03	1	0	1	0	4
	30.03	1	0	1	0	5
	30.03	1	0	1	0	7
	30.03	1	0	1	0	8
	30.03	1	0	1	0	9
	30.03	1	0	1	0	11
	30.03	1	0	1	0	12
	30.03	1	0	1	0	13
	30.03	1	0	1	0	14
	30.03	1	0	1	0	15
	30.03	1	0	1	0	16
	30.03	1	0	1	0	18
	30.03	1	0	1	0	19
	30.03	1	0	1	0	20
	30.03	1	0	1	0	21
	30.03	1	0	1	0	22
	30.03	1	0	1	0	23
	30.03	1	0	1	0	24
	30.03	1	0	1	0	25
	30.03	1	0	1	0	26
Furosemide	330.75	7	4	1	0	3
	330.75	7	4	1	0	4
	330.75	7	4	1	0	5
	330.75	7	4	1	0	7
	330.75	7	4	1	0	9
	330.75	7	4	1	0	10
	330.75	7	4	1	0	11
	330.75	7	4	1	0	12
	330.75	7	4	1	0	13
	330.75	7	4	1	0	14
	330.75	7	4	1	0	15
	330.75	7	4	1	0	16
	330.75	7	4	1	0	17
	330.75	7	4	1	0	19
	330.75	7	4	1	0	21
	330.75	7	4	1	0	22
	330.75	7	4	1	0	23
	330.75	7	4	1	0	25

Gabapentin	30.03	1	0	1	0	27
	171.24	3	2	1	0	3
	171.24	3	2	1	0	4
	171.24	3	2	1	0	5
	171.24	3	2	1	0	7
	171.24	3	2	1	0	8
	171.24	3	2	1	0	9
	171.24	3	2	1	0	12
	171.24	3	2	1	0	13
	171.24	3	2	1	0	15
	171.24	3	2	1	0	16
	171.24	3	2	1	0	17
	171.24	3	2	1	0	21
	171.24	3	2	1	0	23
	171.24	3	2	1	0	24
Glycerol	92.09	3	3	9	0	1
	92.09	3	3	1	0	3
	92.09	3	3	1	0	7
	92.09	3	3	1	0	8
	92.09	3	3	1	0	9
	92.09	3	3	1	0	12
	92.09	3	3	1	0	14
	92.09	3	3	1	0	16
	92.09	3	3	1	0	17
	92.09	3	3	1	0	21
Hydrobromic Acid	171.24	3	2	1	0	25
	80.91	0	0	1	0	16

Hydrochloric Acid	36.46	0	1	1	0	2
	36.46	0	1	1	0	3
	36.46	0	1	1	0	4
	36.46	0	1	1	0	5
	36.46	0	1	1	0	7
	36.46	0	1	1	0	9
	36.46	0	1	1	0	11
	36.46	0	1	1	0	12
	36.46	0	1	1	0	13
	36.46	0	1	1	0	14
	36.46	0	1	1	0	15
	36.46	0	1	1	0	16
	36.46	0	1	1	0	17
	36.46	0	1	1	0	18
	36.46	0	1	1	0	19
	36.46	0	1	1	0	20
	36.46	0	1	1	0	21
	36.46	0	1	1	0	22
	36.46	0	1	1	0	23
	36.46	0	1	1	0	24
	36.46	0	1	1	0	25
	36.46	0	1	1	0	27
Hydrochlorothiazide	297.74	7	4	1	0	3
	297.74	7	4	1	0	4
	297.74	7	4	1	0	5
	297.74	7	4	1	0	7
	297.74	7	4	1	0	9
	297.74	7	4	1	0	11
	297.74	7	4	1	0	12
	297.74	7	4	1	0	13
	297.74	7	4	1	0	14
	297.74	7	4	1	0	16
	297.74	7	4	1	0	17
	297.74	7	4	1	0	21
	297.74	7	4	1	0	22
	297.74	7	4	1	0	23
	297.74	7	4	1	0	24
	297.74	7	4	1	0	25

Hydrofluoric Acid	20.01	1	1	1	0	5
	20.01	1	1	1	0	9
	20.01	1	1	1	0	12
	20.01	1	1	1	0	13
	20.01	1	1	1	0	14
	20.01	1	1	1	0	16
	20.01	1	1	1	0	17
	20.01	1	1	1	0	21
	20.01	1	1	1	0	22
	20.01	1	1	1	0	24
	20.01	1	1	1	0	25
	20.01	1	1	1	0	27
Ibuprofen	206.29	2	1	1	0	2
	206.29	2	1	1	0	3
	206.29	2	1	1	0	4
	206.29	2	1	1	0	5
	206.29	2	1	1	0	7
	206.29	2	1	1	0	8
	206.29	2	1	1	0	9
	206.29	2	1	1	0	10
	206.29	2	1	1	0	11
	206.29	2	1	1	0	12
	206.29	2	1	1	0	13
	206.29	2	1	1	0	14
	206.29	2	1	1	0	15
	206.29	2	1	1	0	16
	206.29	2	1	1	0	17
	206.29	2	1	1	0	19
	206.29	2	1	1	0	21
	206.29	2	1	1	0	22
	206.29	2	1	1	0	23
	206.29	2	1	1	0	24
	206.29	2	1	1	0	25
	206.29	2	1	1	0	26

Imipramine	280.42	2	0	1	0	3
	280.42	2	0	1	0	4
	280.42	2	0	1	0	5
	280.42	2	0	1	0	7
	280.42	2	0	1	0	8
	280.42	2	0	1	0	9
	280.42	2	0	1	0	11
	280.42	2	0	1	0	12
	280.42	2	0	1	0	13
	280.42	2	0	1	0	14
	280.42	2	0	1	0	15
	280.42	2	0	1	0	16
	280.42	2	0	1	0	17
	280.42	2	0	1	0	19
	280.42	2	0	1	0	21
Isopropyl Alcohol	280.42	2	0	1	0	22
	280.42	2	0	1	0	23
	280.42	2	0	1	0	24
	280.42	2	0	1	0	25
Itraconazole	60.10	1	1	1	0	2
	705.65	12	0	1	0	2
	705.65	12	0	1	0	3
	705.65	12	0	1	0	4
	705.65	12	0	1	0	5
	705.65	12	0	1	0	7
	705.65	12	0	1	0	9
	705.65	12	0	1	0	11
	705.65	12	0	1	0	12
	705.65	12	0	1	0	13
	705.65	12	0	1	0	14
	705.65	12	0	1	0	15
	705.65	12	0	1	0	16
	705.65	12	0	1	0	17
	705.65	12	0	1	0	19
	705.65	12	0	1	0	20
	705.65	12	0	1	0	21
	705.65	12	0	1	0	22
	705.65	12	0	1	0	23
	705.65	12	0	1	0	25

Ketoconazole	380.92	1	0	9	0	1
	380.92	1	0	1	0	2
	380.92	1	0	1	0	3
	380.92	1	0	1	0	4
	380.92	1	0	1	0	5
	380.92	1	0	1	0	7
	380.92	1	0	1	0	8
	380.92	1	0	1	0	9
	380.92	1	0	1	0	11
	380.92	1	0	1	0	12
	380.92	1	0	1	0	13
	380.92	1	0	1	0	14
	380.92	1	0	1	0	15
	380.92	1	0	1	0	16
	380.92	1	0	1	0	17
	380.92	1	0	1	0	20
	380.92	1	0	1	0	21
	380.92	1	0	1	0	22
	380.92	1	0	1	0	23
	380.92	1	0	1	0	24
	380.92	1	0	1	0	25
	380.92	1	0	1	0	27
Ketoprofen	254.29	3	1	1	0	3
	254.29	3	1	1	0	4
	254.29	3	1	1	0	5
	254.29	3	1	1	0	9
	254.29	3	1	1	0	11
	254.29	3	1	1	0	12
	254.29	3	1	1	0	13
	254.29	3	1	1	0	14
	254.29	3	1	1	0	16
	254.29	3	1	1	0	21
	254.29	3	1	1	0	22
	254.29	3	1	1	0	23
	254.29	3	1	1	0	25
	254.29	3	1	1	0	27

Labetalol-HCl	328.42	5	5	1	0	2
	328.42	5	5	1	0	3
	328.42	5	5	1	0	4
	328.42	5	5	1	0	4
	328.42	5	5	1	0	5
	328.42	5	5	1	0	7
	328.42	5	5	1	0	9
	328.42	5	5	1	0	11
	328.42	5	5	1	0	12
	328.42	5	5	1	0	13
	328.42	5	5	1	0	14
	328.42	5	5	1	0	15
	328.42	5	5	1	0	16
	328.42	5	5	1	0	17
	328.42	5	5	1	0	21
	328.42	5	5	1	0	22
Lisinopril	328.42	5	5	1	0	23
	328.42	5	5	1	0	25
	405.50	8	5	1	0	3
	405.50	8	5	1	0	4
	405.50	8	5	1	0	5
	405.50	8	5	1	0	7
	405.50	8	5	1	0	8
	405.50	8	5	1	0	9
	405.50	8	5	1	0	11
	405.50	8	5	1	0	12
	405.50	8	5	1	0	13
	405.50	8	5	1	0	14
	405.50	8	5	1	0	16
	405.50	8	5	1	0	17
	405.50	8	5	1	0	21
	405.50	8	5	1	0	22
	405.50	8	5	1	0	23
	405.50	8	5	1	0	24
	405.50	8	5	1	0	25

Magnesium Sulfate	120.37	4	0	1	0	2
	120.37	4	0	1	0	3
	120.37	4	0	1	0	4
	120.37	4	0	1	0	5
	120.37	4	0	1	0	8
	120.37	4	0	1	0	9
	120.37	4	0	1	0	11
	120.37	4	0	1	0	12
	120.37	4	0	1	0	13
	120.37	4	0	1	0	14
	120.37	4	0	1	0	16
	120.37	4	0	1	0	17
	120.37	4	0	1	0	21
	120.37	4	0	1	0	22
	120.37	4	0	1	0	25
	120.37	4	0	1	0	27
Mannitol	182.18	6	6	1	0	3
	182.18	6	6	1	0	5
	182.18	6	6	1	0	7
	182.18	6	6	1	0	8
	182.18	6	6	1	0	9
	182.18	6	6	1	0	11
	182.18	6	6	1	0	12
	182.18	6	6	1	0	14
	182.18	6	6	1	0	15
	182.18	6	6	1	0	16
	182.18	6	6	1	0	17
	182.18	6	6	1	0	21
	182.18	6	6	1	0	22
	182.18	6	6	1	0	23
	182.18	6	6	1	0	24
	182.18	6	6	1	0	25
	182.18	6	6	1	0	26
	182.18	6	6	1	0	27

Methotrexate	454.45	13	7	1	0	2
	454.45	13	7	1	0	3
	454.45	13	7	1	0	4
	454.45	13	7	1	0	4
	454.45	13	7	1	0	5
	454.45	13	7	1	0	7
	454.45	13	7	1	0	8
	454.45	13	7	1	0	9
	454.45	13	7	1	0	10
	454.45	13	7	1	0	11
	454.45	13	7	1	0	12
	454.45	13	7	1	0	13
	454.45	13	7	1	0	14
	454.45	13	7	1	0	15
	454.45	13	7	1	0	16
	454.45	13	7	1	0	17
	454.45	13	7	1	0	19
	454.45	13	7	1	0	21
	454.45	13	7	1	0	22
	454.45	13	7	1	0	23
	454.45	13	7	1	0	24
	454.45	13	7	1	0	25
	454.45	13	7	1	0	26
	454.45	13	7	1	0	27
Metoprolol-tartrate	267.37	4	2	1	0	3
	267.37	4	2	1	0	4
	267.37	4	2	1	0	5
	267.37	4	2	1	0	7
	267.37	4	2	1	0	8
	267.37	4	2	1	0	9
	267.37	4	2	1	0	11
	267.37	4	2	1	0	12
	267.37	4	2	1	0	13
	267.37	4	2	1	0	14
	267.37	4	2	1	0	15
	267.37	4	2	1	0	16
	267.37	4	2	1	0	17
	267.37	4	2	1	0	21
	267.37	4	2	1	0	22
	267.37	4	2	1	0	23
	267.37	4	2	1	0	24
	267.37	4	2	1	0	25
	267.37	4	2	1	0	27

Nadolol	309.41	5	4	1	0	3
	309.41	5	4	1	0	5
	309.41	5	4	1	0	7
	309.41	5	4	1	0	11
	309.41	5	4	1	0	13
	309.41	5	4	1	0	16
	309.41	5	4	1	0	21
	309.41	5	4	1	0	23
	309.41	5	4	1	0	24
	309.41	5	4	1	0	25
Naloxone	327.38	5	2	1	0	2
	327.38	5	2	1	0	3
	327.38	5	2	1	0	4
	327.38	5	2	1	0	5
	327.38	5	2	1	0	7
	327.38	5	2	1	0	8
	327.38	5	2	1	0	9
	327.38	5	2	1	0	10
	327.38	5	2	1	0	11
	327.38	5	2	1	0	12
	327.38	5	2	1	0	13
	327.38	5	2	1	0	14
	327.38	5	2	1	0	15
	327.38	5	2	1	0	16
	327.38	5	2	1	0	17
	327.38	5	2	1	0	19
	327.38	5	2	1	0	21
	327.38	5	2	1	0	22
	327.38	5	2	1	0	23
	327.38	5	2	1	0	24
	327.38	5	2	1	0	25
	327.38	5	2	1	0	27

Naproxen-sodium	230.27	3	1	1	0	2
	230.27	3	1	1	0	3
	230.27	3	1	1	0	4
	230.27	3	1	1	0	5
	230.27	3	1	1	0	7
	230.27	3	1	1	0	9
	230.27	3	1	1	0	11
	230.27	3	1	1	0	12
	230.27	3	1	1	0	13
	230.27	3	1	1	0	14
	230.27	3	1	1	0	16
	230.27	3	1	1	0	17
	230.27	3	1	1	0	19
	230.27	3	1	1	0	21
	230.27	3	1	1	0	22
	230.27	3	1	1	0	23
	230.27	3	1	1	0	24
	230.27	3	1	1	0	25
	230.27	3	1	1	0	26
	230.27	3	1	1	0	27
Nortriptylene-HCl	263.39	1	1	1	0	3
	263.39	1	1	1	0	4
	263.39	1	1	1	0	5
	263.39	1	1	1	0	8
	263.39	1	1	1	0	9
	263.39	1	1	1	0	12
	263.39	1	1	1	0	13
	263.39	1	1	1	0	16
	263.39	1	1	1	0	17
	263.39	1	1	1	0	19
	263.39	1	1	1	0	21
	263.39	1	1	1	0	22
	263.39	1	1	1	0	24
	263.39	1	1	1	0	25
	263.39	1	1	1	0	27

Omeprazole	267.25	9	2	1	0	2
	267.25	9	2	1	0	3
	267.25	9	2	1	0	4
	267.25	9	2	1	0	5
	267.25	9	2	1	0	7
	267.25	9	2	1	0	8
	267.25	9	2	1	0	9
	267.25	9	2	1	0	11
	267.25	9	2	1	0	12
	267.25	9	2	1	0	13
	267.25	9	2	1	0	14
	267.25	9	2	1	0	15
	267.25	9	2	1	0	16
	267.25	9	2	1	0	17
	267.25	9	2	1	0	19
	267.25	9	2	1	0	20
	267.25	9	2	1	0	21
	267.25	9	2	1	0	22
	267.25	9	2	1	0	23
	267.25	9	2	1	0	24
	267.25	9	2	1	0	25
Phenytoin	451.49	10	2	1	0	3
	451.49	10	2	1	0	4
	451.49	10	2	1	0	5
	451.49	10	2	1	0	7
	451.49	10	2	1	0	8
	451.49	10	2	1	0	9
	451.49	10	2	1	0	10
	451.49	10	2	1	0	11
	451.49	10	2	1	0	12
	451.49	10	2	1	0	13
	451.49	10	2	1	0	14
	451.49	10	2	1	0	15
	451.49	10	2	1	0	16
	451.49	10	2	1	0	17
	451.49	10	2	1	0	19
	451.49	10	2	1	0	21
	451.49	10	2	1	0	22
	451.49	10	2	1	0	23
	451.49	10	2	1	0	24
	451.49	10	2	1	0	25
	451.49	10	2	1	0	26
	451.49	10	2	1	0	27

Piroxicam	331.35	7	2	1	0	3
	331.35	7	2	1	0	4
	331.35	7	2	1	0	5
	331.35	7	2	1	0	8
	331.35	7	2	1	0	9
	331.35	7	2	1	0	11
	331.35	7	2	1	0	12
	331.35	7	2	1	0	14
	331.35	7	2	1	0	15
	331.35	7	2	1	0	16
	331.35	7	2	1	0	17
	331.35	7	2	1	0	19
	331.35	7	2	1	0	21
	331.35	7	2	1	0	22
	331.35	7	2	1	0	23
	331.35	7	2	1	0	25
	331.35	7	2	1	0	26
	331.35	7	2	1	0	27
Potassium Bromide	119.00	1	0	1	0	16
Potassium Permanganate	158.03	4	0	1	0	5
Prazosin	383.41	9	2	1	0	3
	383.41	9	2	1	0	4
	383.41	9	2	1	0	5
	383.41	9	2	1	0	7
	383.41	9	2	1	0	8
	383.41	9	2	1	0	9
	383.41	9	2	1	0	11
	383.41	9	2	1	0	12
	383.41	9	2	1	0	13
	383.41	9	2	1	0	14
	383.41	9	2	1	0	15
	383.41	9	2	1	0	16
	383.41	9	2	1	0	17
	383.41	9	2	1	0	21
	383.41	9	2	1	0	23
	383.41	9	2	1	0	25

Propranolol-HCl	259.35	3	2	1	0	3
	259.35	3	2	1	0	4
	259.35	3	2	1	0	5
	259.35	3	2	1	0	7
	259.35	3	2	1	0	8
	259.35	3	2	1	0	9
	259.35	3	2	1	0	11
	259.35	3	2	1	0	12
	259.35	3	2	1	0	13
	259.35	3	2	1	0	14
	259.35	3	2	1	0	15
	259.35	3	2	1	0	16
	259.35	3	2	1	0	17
	259.35	3	2	1	0	19
	259.35	3	2	1	0	21
	259.35	3	2	1	0	22
	259.35	3	2	1	0	23
	259.35	3	2	1	0	24
	259.35	3	2	1	0	25
	259.35	3	2	1	0	27
Quinidine	324.43	4	1	1	0	3
	324.43	4	1	1	0	4
	324.43	4	1	1	0	5
	324.43	4	1	1	0	8
	324.43	4	1	1	0	9
	324.43	4	1	1	0	10
	324.43	4	1	1	0	11
	324.43	4	1	1	0	12
	324.43	4	1	1	0	13
	324.43	4	1	1	0	14
	324.43	4	1	1	0	16
	324.43	4	1	1	0	17
	324.43	4	1	1	0	19
	324.43	4	1	1	0	20
	324.43	4	1	1	0	21
	324.43	4	1	1	0	22
	324.43	4	1	1	0	23
	324.43	4	1	1	0	25

Ranitidine-HCl	314.41	7	2	1	0	2
	314.41	7	2	1	0	3
	314.41	7	2	1	0	3
	314.41	7	2	1	0	5
	314.41	7	2	1	0	7
	314.41	7	2	1	0	8
	314.41	7	2	1	0	9
	314.41	7	2	1	0	11
	314.41	7	2	1	0	12
	314.41	7	2	1	0	13
	314.41	7	2	1	0	14
	314.41	7	2	1	0	15
	314.41	7	2	1	0	16
	314.41	7	2	1	0	17
	314.41	7	2	1	0	19
	314.41	7	2	1	0	21
	314.41	7	2	1	0	22
	314.41	7	2	1	0	23
	314.41	7	2	1	0	24
	314.41	7	2	1	0	25
	314.41	7	2	1	0	26
	314.41	7	2	1	0	27
Silver Nitrate	169.87	3	0	1	0	11
Sodium Thiosulfate	158.11	4	0	1	0	9
	158.11	4	0	1	0	12
	158.11	4	0	1	0	16
	158.11	4	0	1	0	17
	158.11	4	0	1	0	19
	158.11	4	0	1	0	25
Tenidap	320.76	5	2	1	0	9
	320.76	5	2	1	0	11
	320.76	5	2	1	0	12
	320.76	5	2	1	0	14
	320.76	5	2	1	0	23

Terfenadine	471.69	3	2	1	0	3
	471.69	3	2	1	0	4
	471.69	3	2	1	0	5
	471.69	3	2	1	0	9
	471.69	3	2	1	0	11
	471.69	3	2	1	0	12
	471.69	3	2	1	0	14
	471.69	3	2	1	0	15
	471.69	3	2	1	0	16
	471.69	3	2	1	0	17
	471.69	3	2	1	0	19
	471.69	3	2	1	0	21
	471.69	3	2	1	0	22
	471.69	3	2	1	0	23
	471.69	3	2	1	0	25
Testosterone	288.43	2	1	9	0	1
	288.43	2	1	1	0	3
	288.43	2	1	1	0	4
	288.43	2	1	1	0	5
	288.43	2	1	1	0	7
	288.43	2	1	1	0	9
	288.43	2	1	1	0	10
	288.43	2	1	1	0	11
	288.43	2	1	1	0	12
	288.43	2	1	1	0	13
	288.43	2	1	1	0	15
	288.43	2	1	1	0	16
	288.43	2	1	1	0	17
	288.43	2	1	1	0	19
	288.43	2	1	1	0	20
	288.43	2	1	1	0	21
	288.43	2	1	1	0	22
	288.43	2	1	1	0	23
	288.43	2	1	1	0	25
	288.43	2	1	1	0	26
	288.43	2	1	1	0	27
Trovafloxacin	416.36	7	3	1	0	2
	416.36	7	3	1	0	3
	416.36	7	3	1	0	5
	416.36	7	3	1	0	9
	416.36	7	3	1	0	12
	416.36	7	3	1	0	14
	416.36	7	3	1	0	16
	416.36	7	3	1	0	21
	416.36	7	3	1	0	25

Valproic-acid	144.22	2	1	9	0	1
	144.22	2	1	1	0	3
	144.22	2	1	1	0	4
	144.22	2	1	1	0	5
	144.22	2	1	1	0	7
	144.22	2	1	1	0	8
	144.22	2	1	1	0	9
	144.22	2	1	1	0	10
	144.22	2	1	1	0	11
	144.22	2	1	1	0	12
	144.22	2	1	1	0	13
	144.22	2	1	1	0	14
	144.22	2	1	1	0	15
	144.22	2	1	1	0	16
	144.22	2	1	1	0	17
	144.22	2	1	1	0	19
	144.22	2	1	1	0	21
	144.22	2	1	1	0	22
	144.22	2	1	1	0	23
	144.22	2	1	1	0	24
	144.22	2	1	1	0	25
	144.22	2	1	1	0	27
Vinblastine	811.00	13	3	1	0	2
	811.00	13	3	1	0	3
	811.00	13	3	1	0	4
	811.00	13	3	1	0	5
	811.00	13	3	1	0	7
	811.00	13	3	1	0	8
	811.00	13	3	1	0	9
	811.00	13	3	1	0	10
	811.00	13	3	1	0	11
	811.00	13	3	1	0	12
	811.00	13	3	1	0	13
	811.00	13	3	1	0	14
	811.00	13	3	1	0	15
	811.00	13	3	1	0	16
	811.00	13	3	1	0	17
	811.00	13	3	1	0	19
	811.00	13	3	1	0	21
	811.00	13	3	1	0	22
	811.00	13	3	1	0	23
	811.00	13	3	1	0	24
	811.00	13	3	1	0	25
	811.00	13	3	1	0	26

Zinc Chloride	136.29	0	0	1	0	5
	136.29	0	0	1	0	13
	136.29	0	0	1	0	15
	136.29	0	0	1	0	21
	136.29	0	0	1	0	25
Ziprasidone	412.95	5	1	1	0	3
	412.95	5	1	1	0	4
	412.95	5	1	1	0	5
	412.95	5	1	1	0	7
	412.95	5	1	1	0	11
	412.95	5	1	1	0	13
	412.95	5	1	1	0	14
	412.95	5	1	1	0	16
	412.95	5	1	1	0	19
	412.95	5	1	1	0	21
	412.95	5	1	1	0	22
	412.95	5	1	1	0	23
	412.95	5	1	1	0	25

Appendix C: Additional MATLAB Training Sessions

Figures C.1-C.7 shows the ANN simulations results for a network using the training function `trainscg`. The `trainscg` training function is similar the default function shown in Chapter 4, but it uses a scaled conjugate gradient to train the data in the network. `Trainscg` also does not use `mu` as a network parameter to measure how well the network is performing during training. This session was terminated due to the upper limits of epochs being reached. Even though the network reached the maximum number of epochs, it still performed fairly well with an R-value of 0.93898. Figure C.2 shows that the actual versus ANN derived disease plot appeared to follow a linear line but did have some curve to it. This curve to the output data would explain the R-value being less than one. The performance and MSE plots (figures C.3 and C.4) look identical because MSE is used with `trainscg` to measure the network performance. It should also be noted that the network was consistently able to lower the performance/MSE over the course of the simulation indicating the network was able to continually improve over time. Figure C.5 shows the regression plot which makes it clear that the ANN derived disease values did not directly lay along the one-to-one slope. Figure C.6 shows numerous fluctuations in the gradient and the network tried to increase the performance and Figure C.7 shows the error histogram highlighting that most of the error in the network was small, but not small enough to produce a linear output.

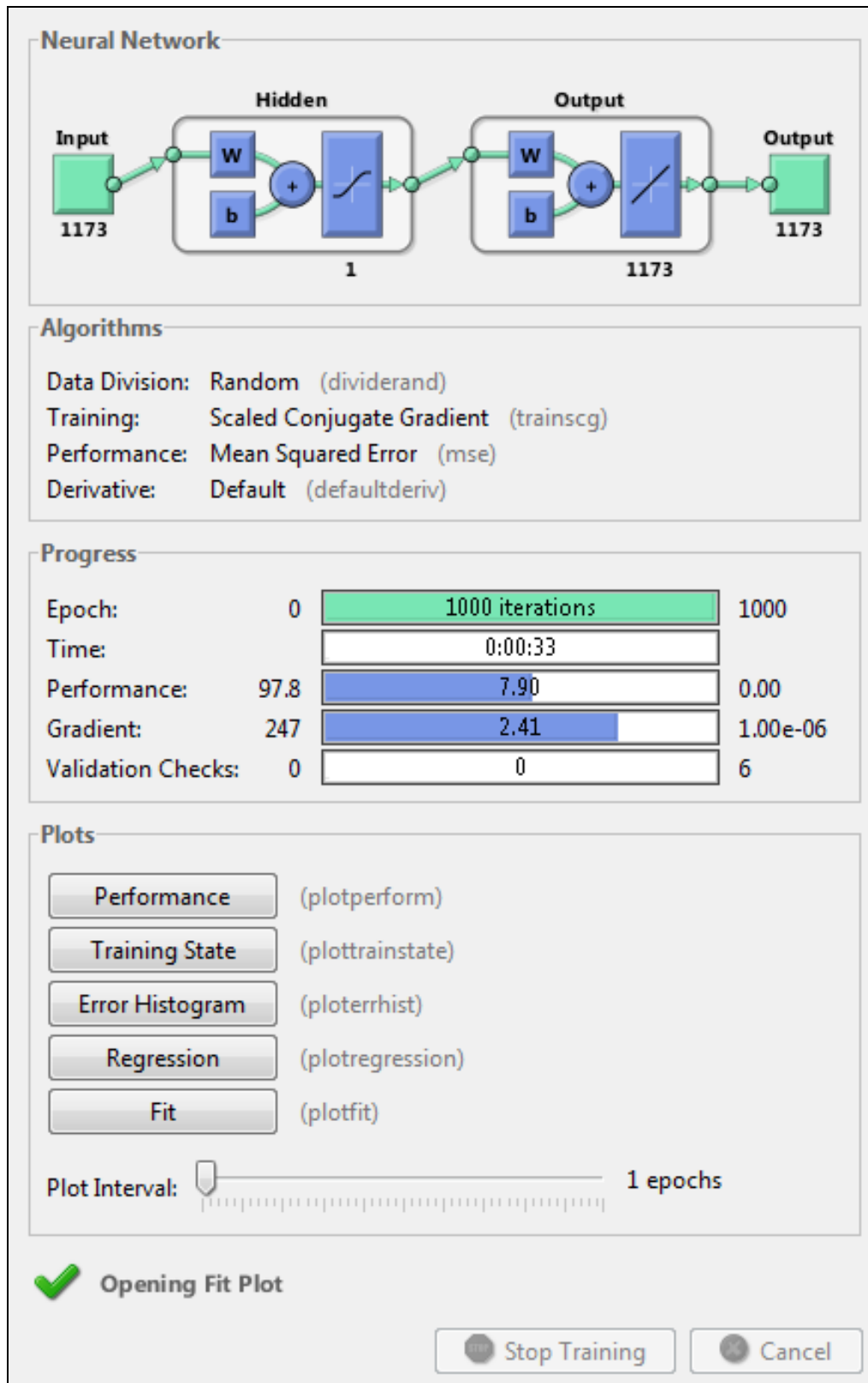


Figure C.1: Trainscg Training Session

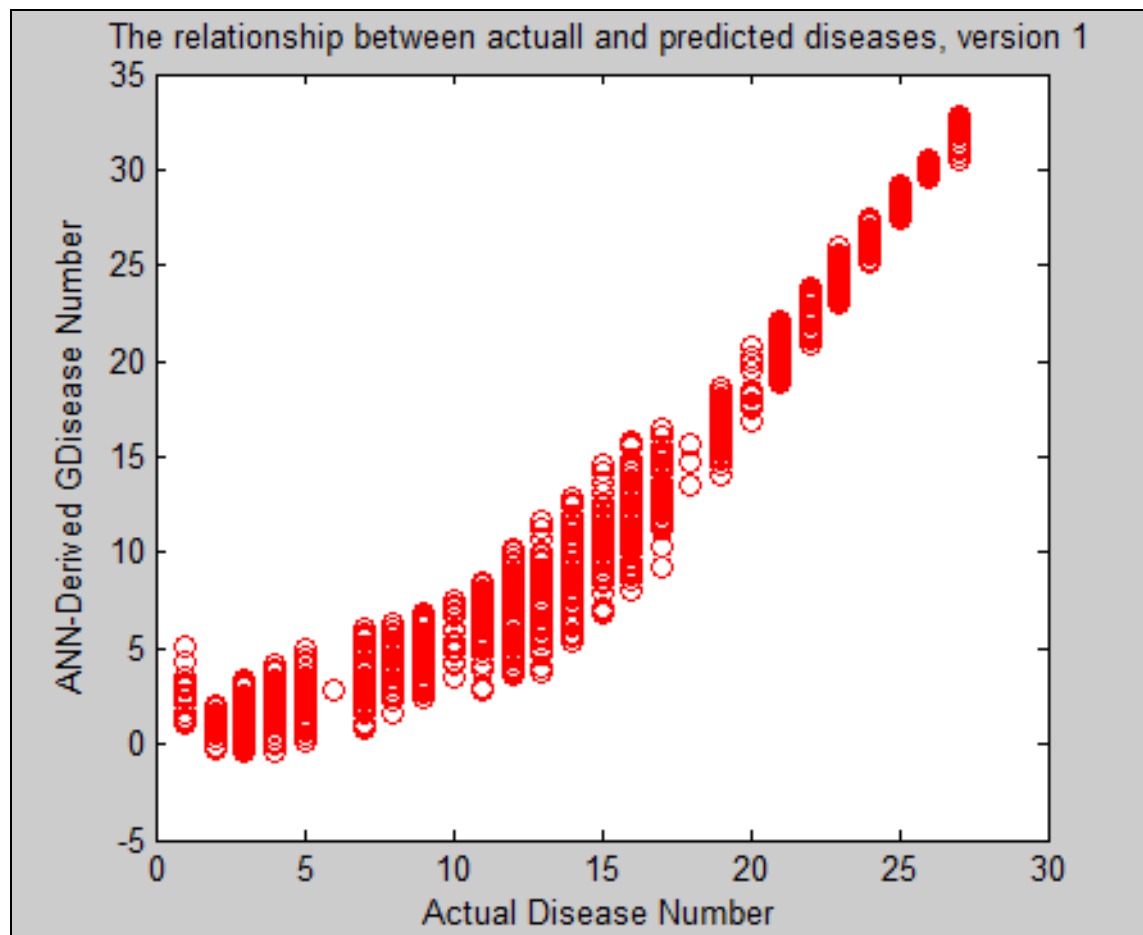


Figure C.2: Trainseg Actual Disease versus ANN Derived Disease Outputs Plot

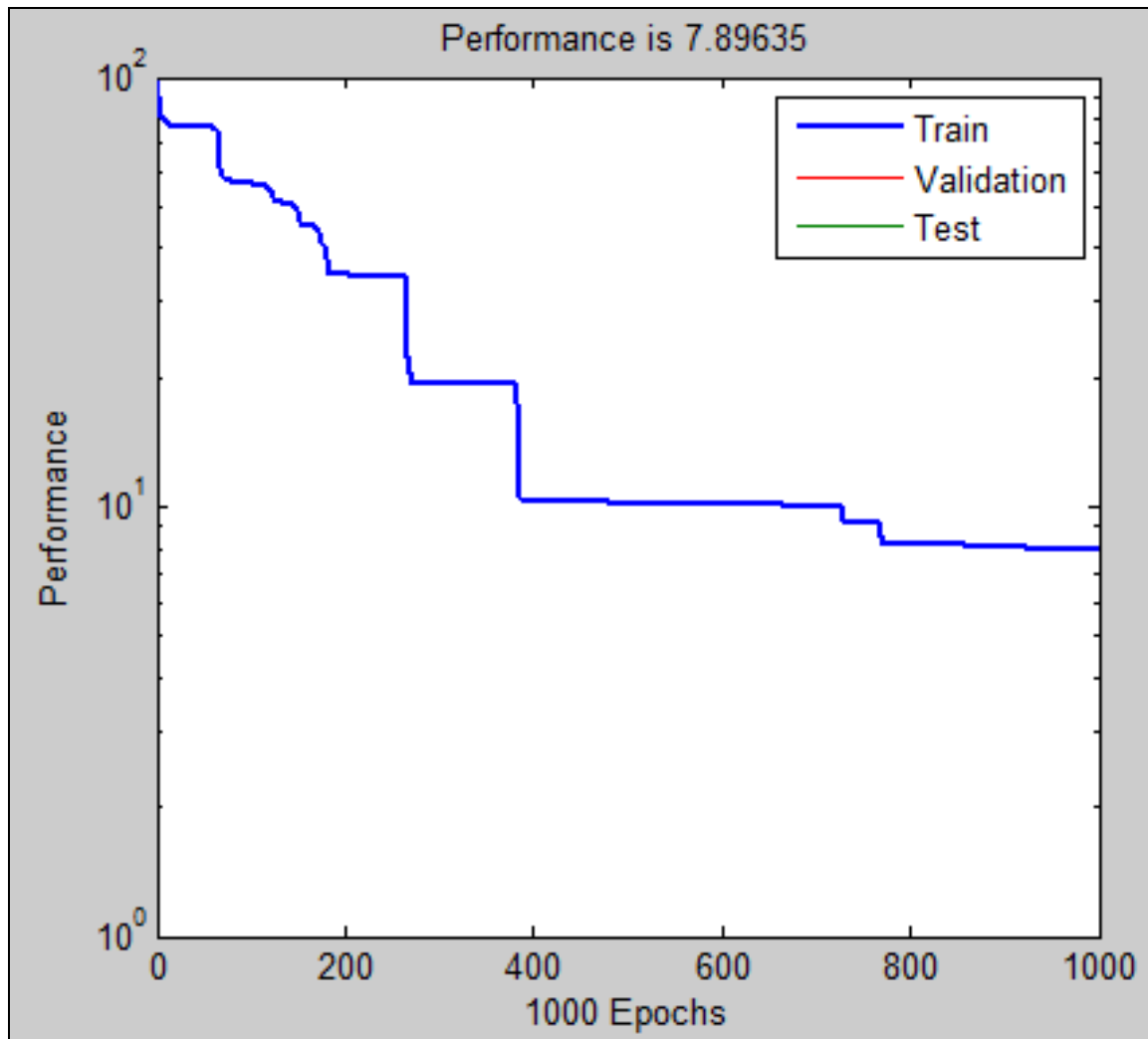


Figure C.3: Trainscg Performance Plot

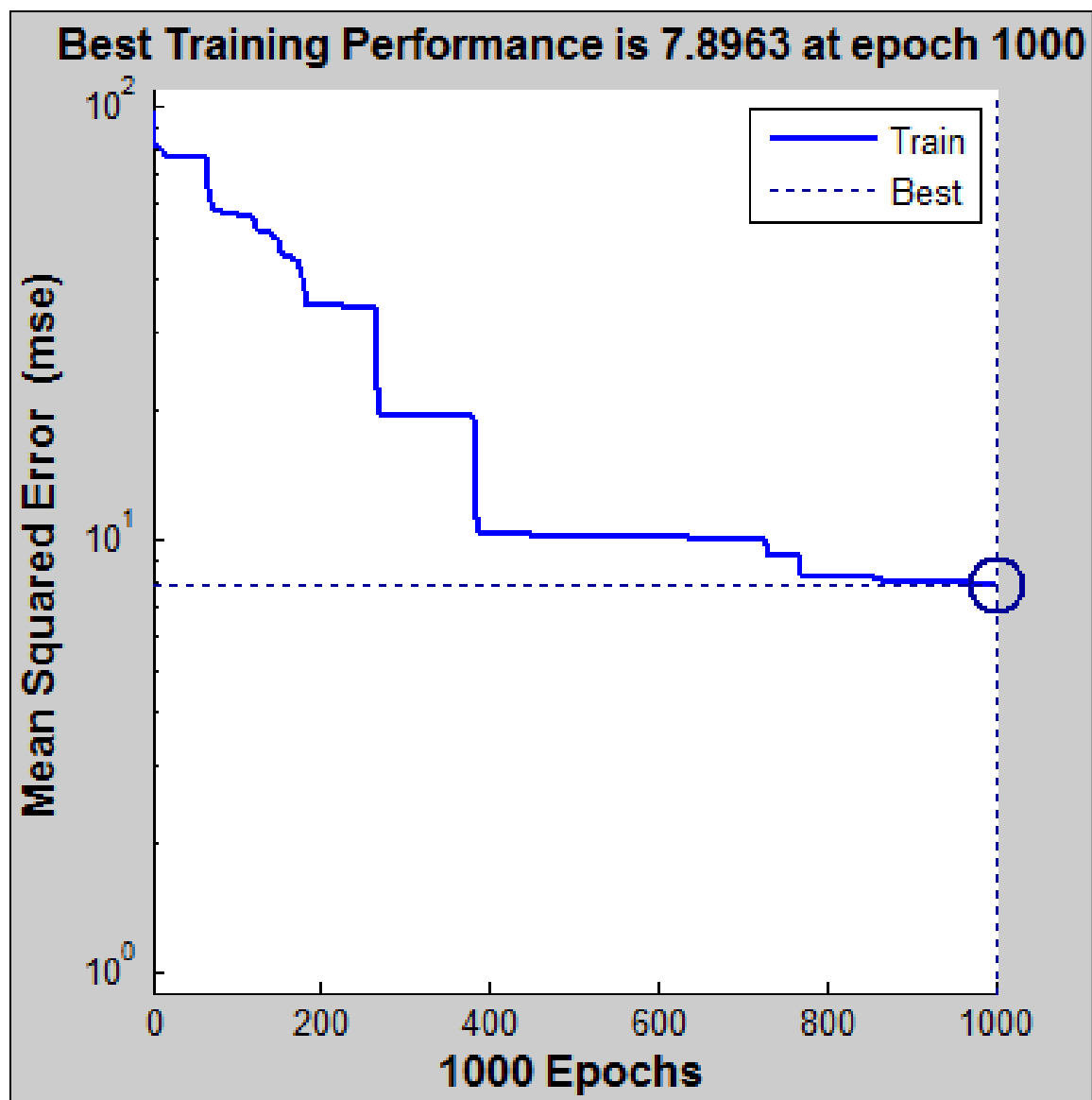


Figure C.4: Trainscg Mean Squared Error Plot

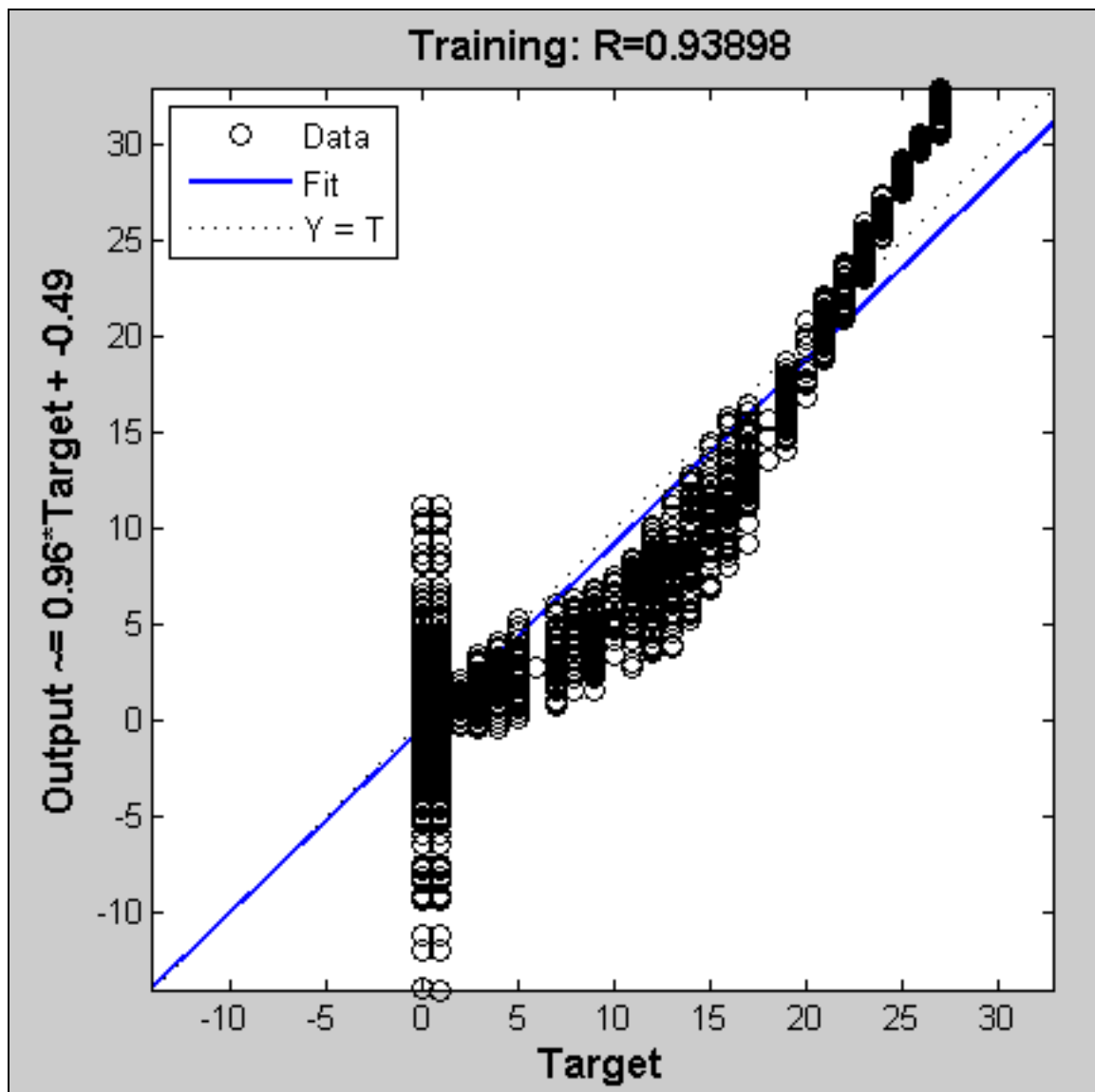


Figure C.5: Trainscg Regression Plot

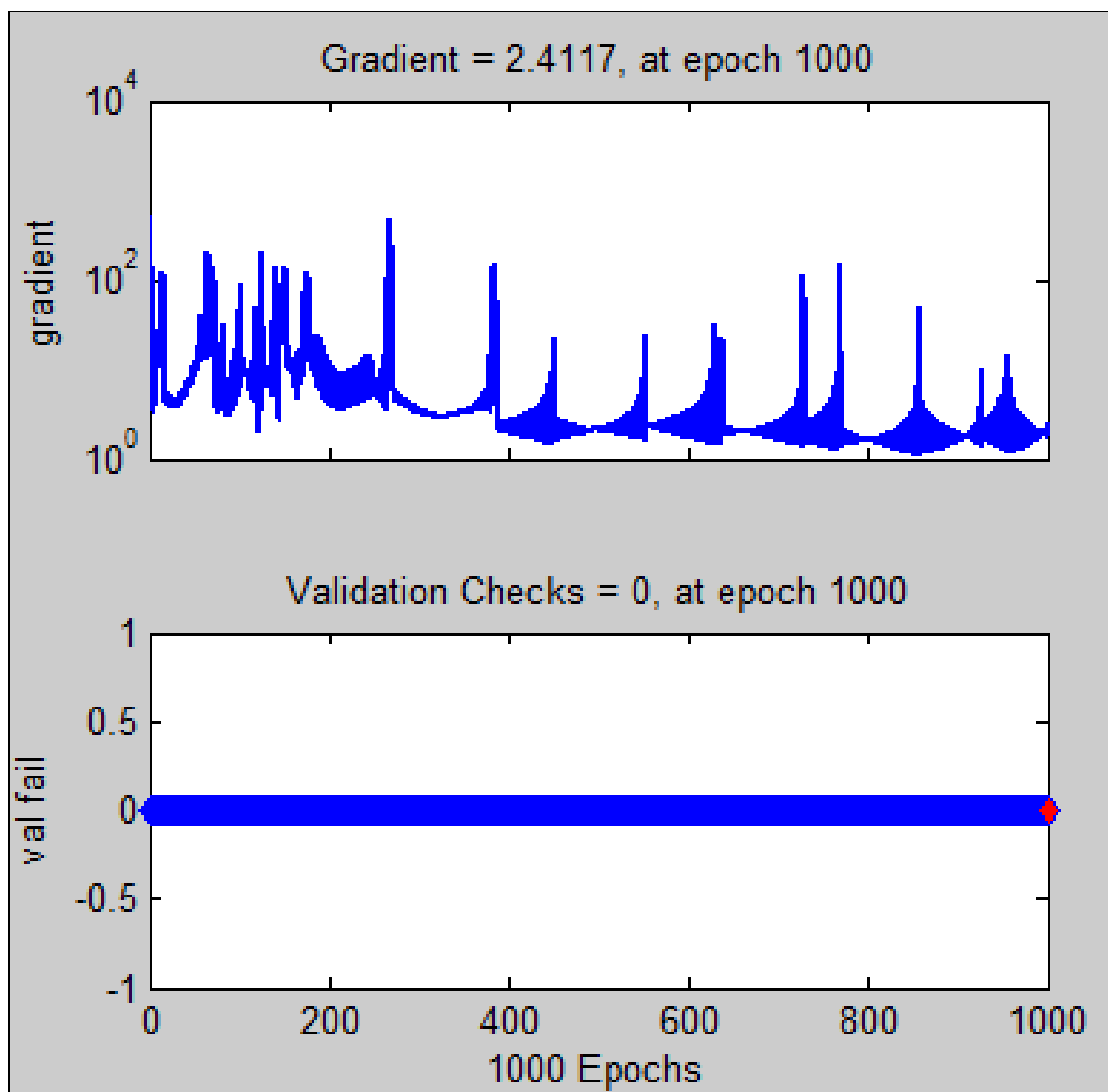


Figure C.6: Transcg Training States

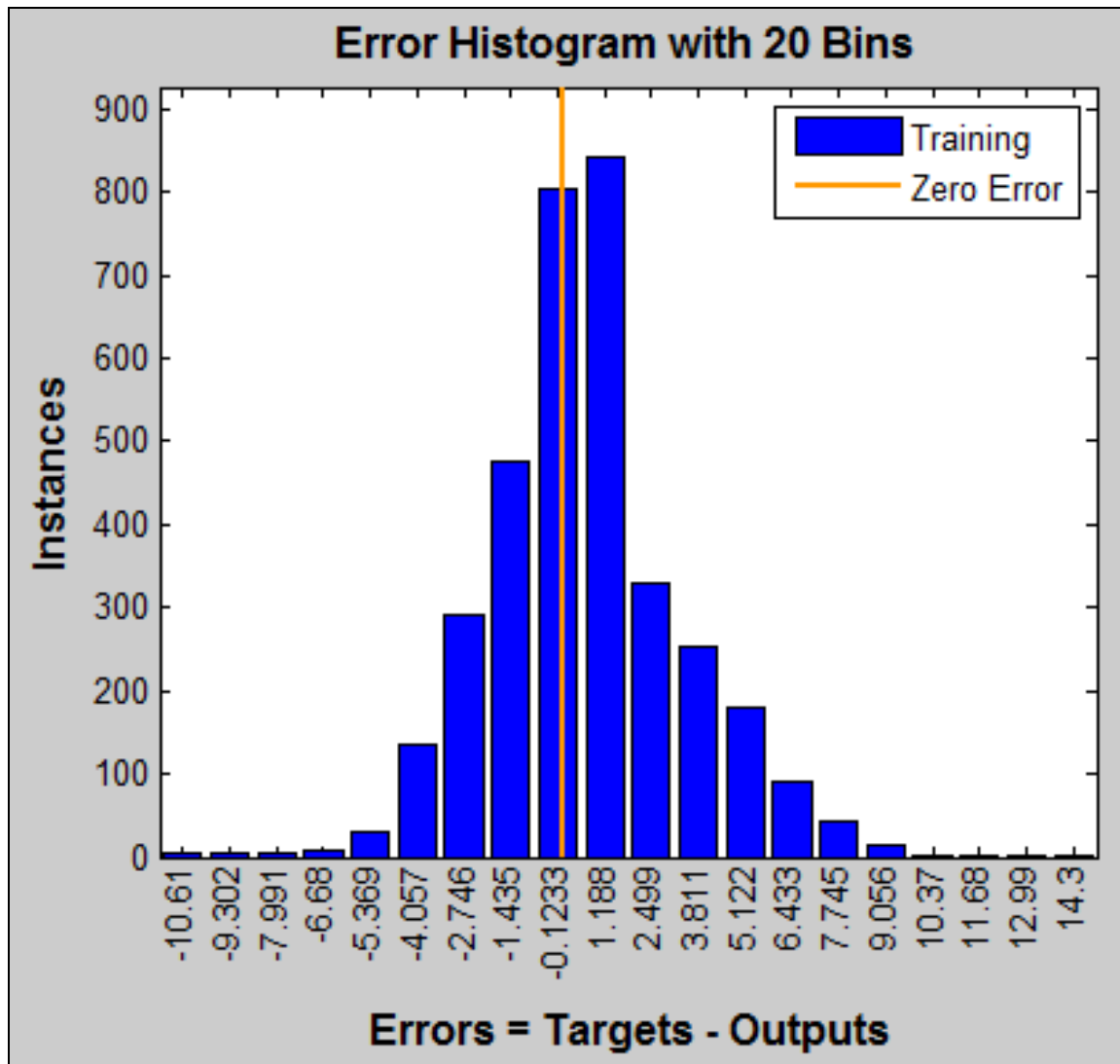


Figure C.7: Trainscg Error Histogram

Figures C.8-C.14 shows the ANN simulations results for a network using the training function `trainrp`. The `trainrp` training function is similar the default function shown in Chapter 4, but it uses resilient backpropagation to train the data in the network. `Trainrp` also does not use `mu` as a network parameter to measure how well the network is performing during training. This session was terminated due to the gradient lower limit indicating the network potentially performed well. However, then reviewing Figure C.9 and the R-value, it is apparent the ANN derived disease values do not follow a one-to-one slope. In addition to the R-value of 0.53903, the regressions plot in figure C.12 shows the ANN derived disease values follow a slope of 0.5 to 1 which indicates the network predict values half that of the actual values. The performance and MSE plots (Figures C.9 and C.10) look identical because MSE is used with `trainscg` to measure the network performance. Figure C.13 shows a consistently decreasing gradient and Figure C.14 shows the error histogram highlighting that minimal error occurred over a wide range, the majority of the error was small.

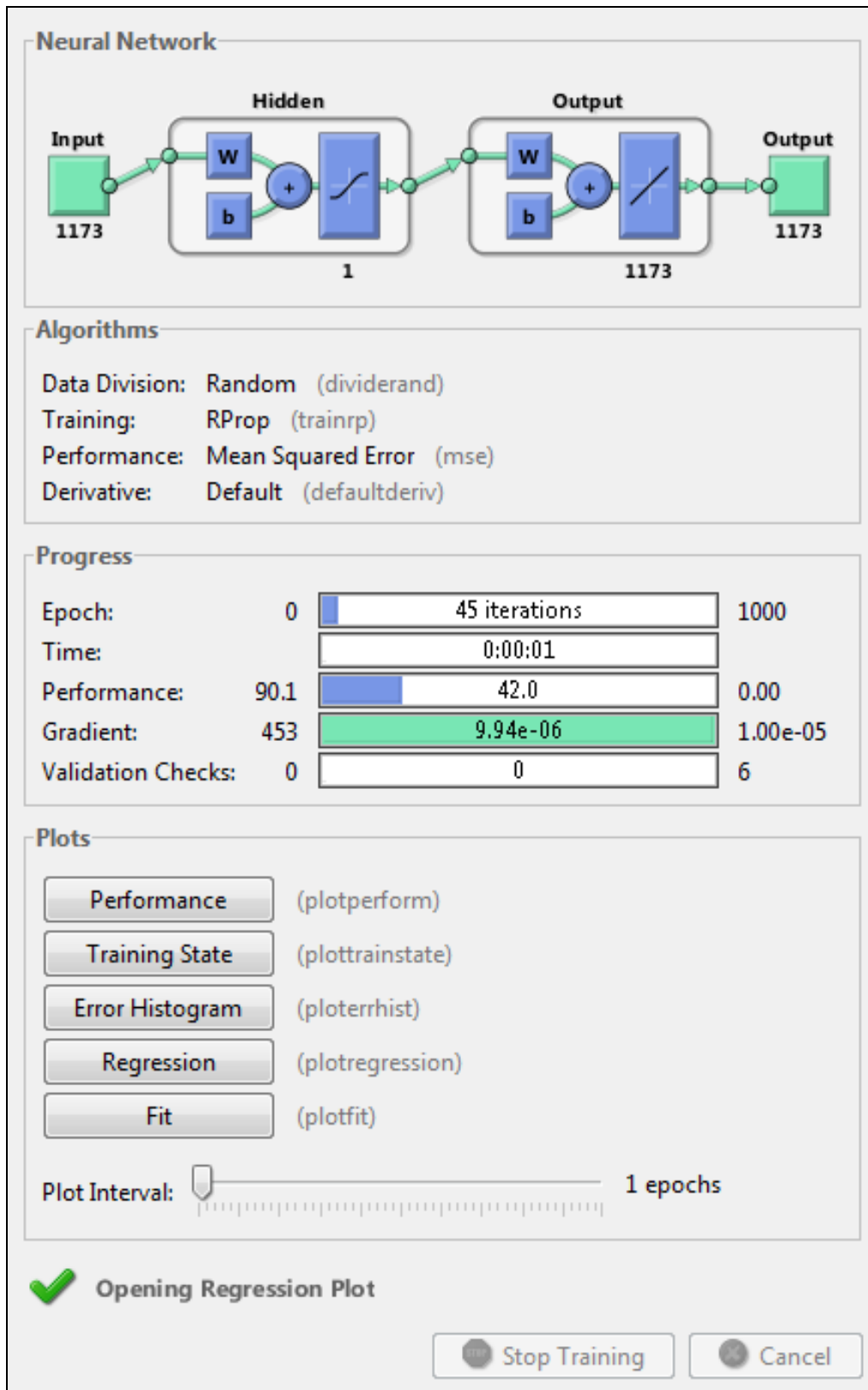


Figure C.8: Trainrp Training Session

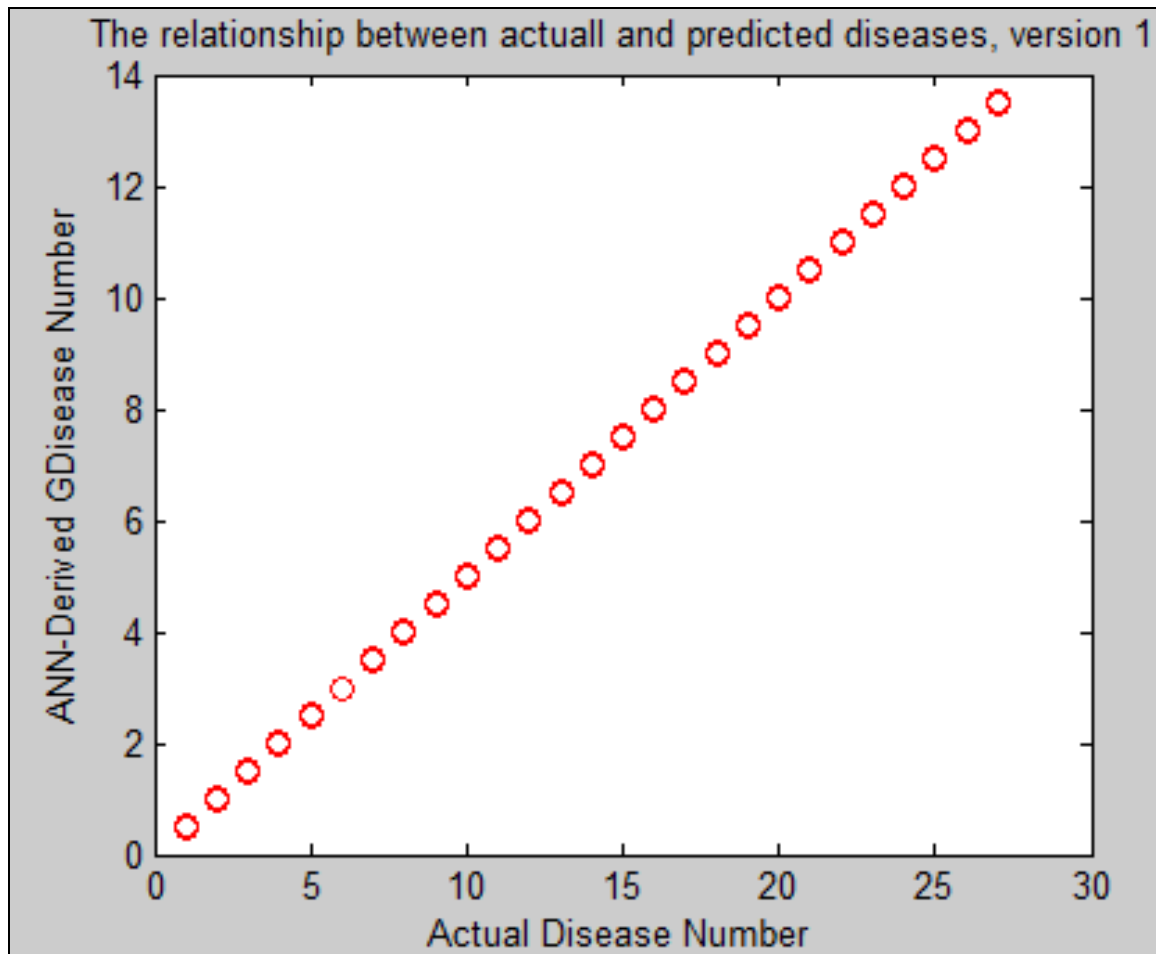


Figure C.9: Trainsrp Actual Disease versus ANN Derived Disease Outputs Plot

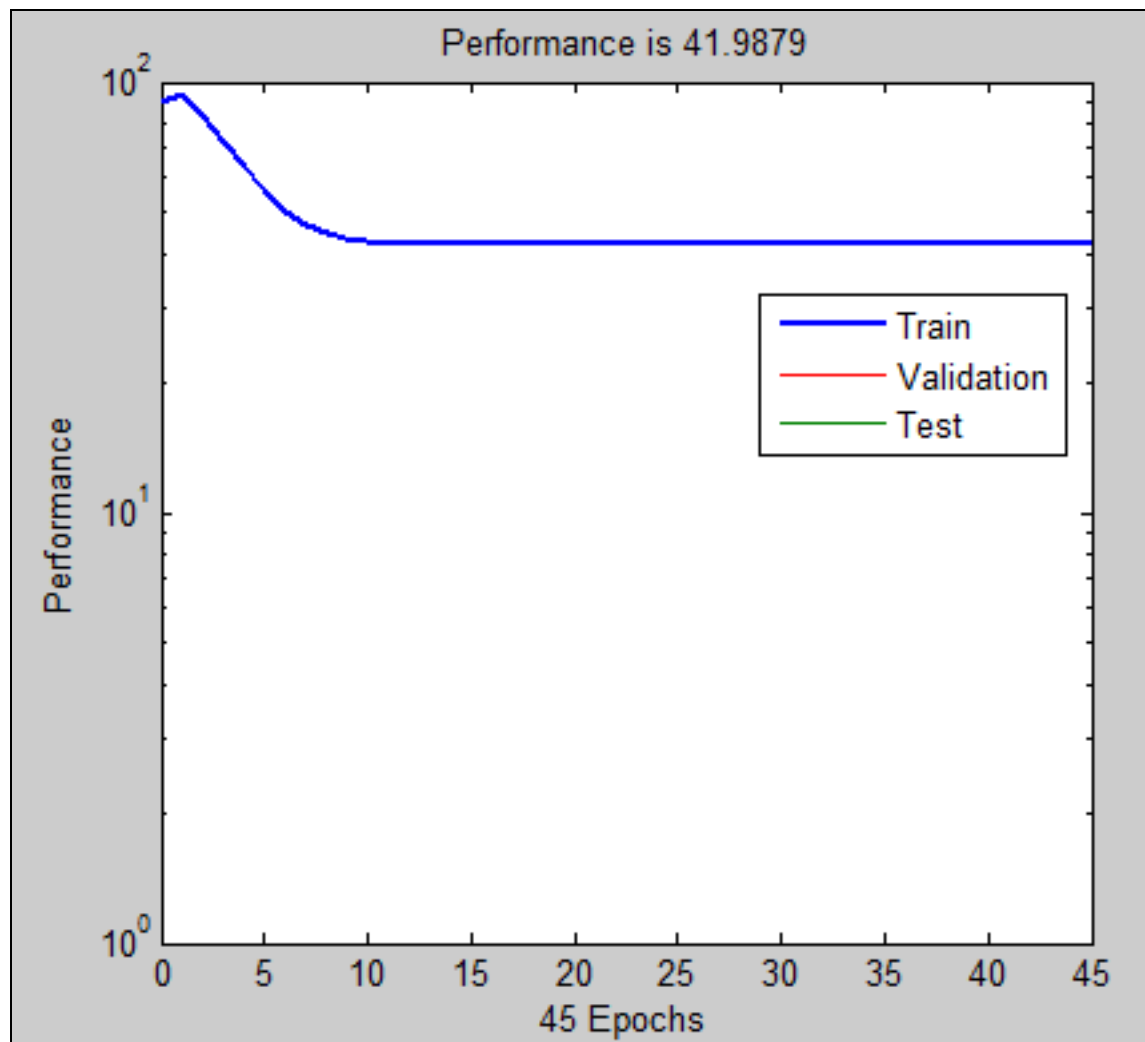


Figure C.10: Trainrp Performance Plot

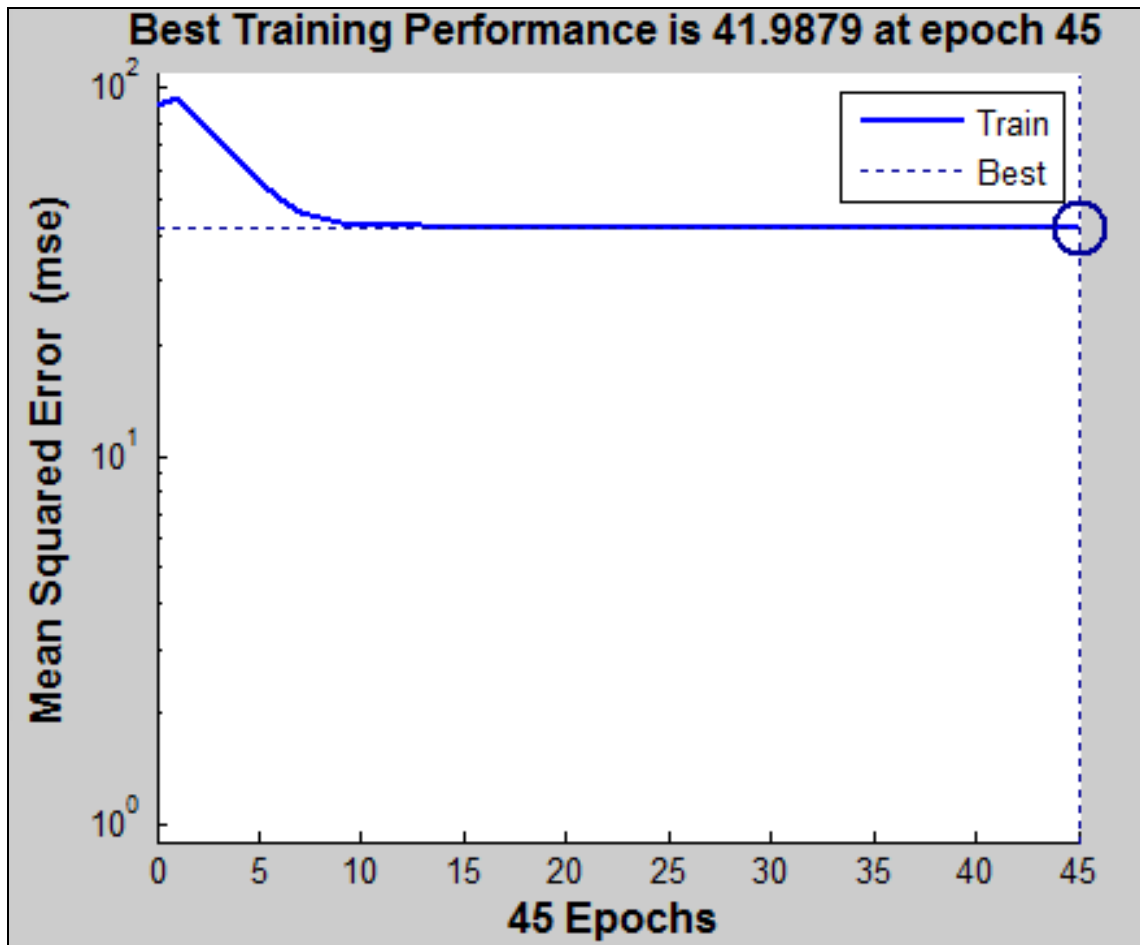


Figure C.11: Trainrp Mean Squared Error Plot

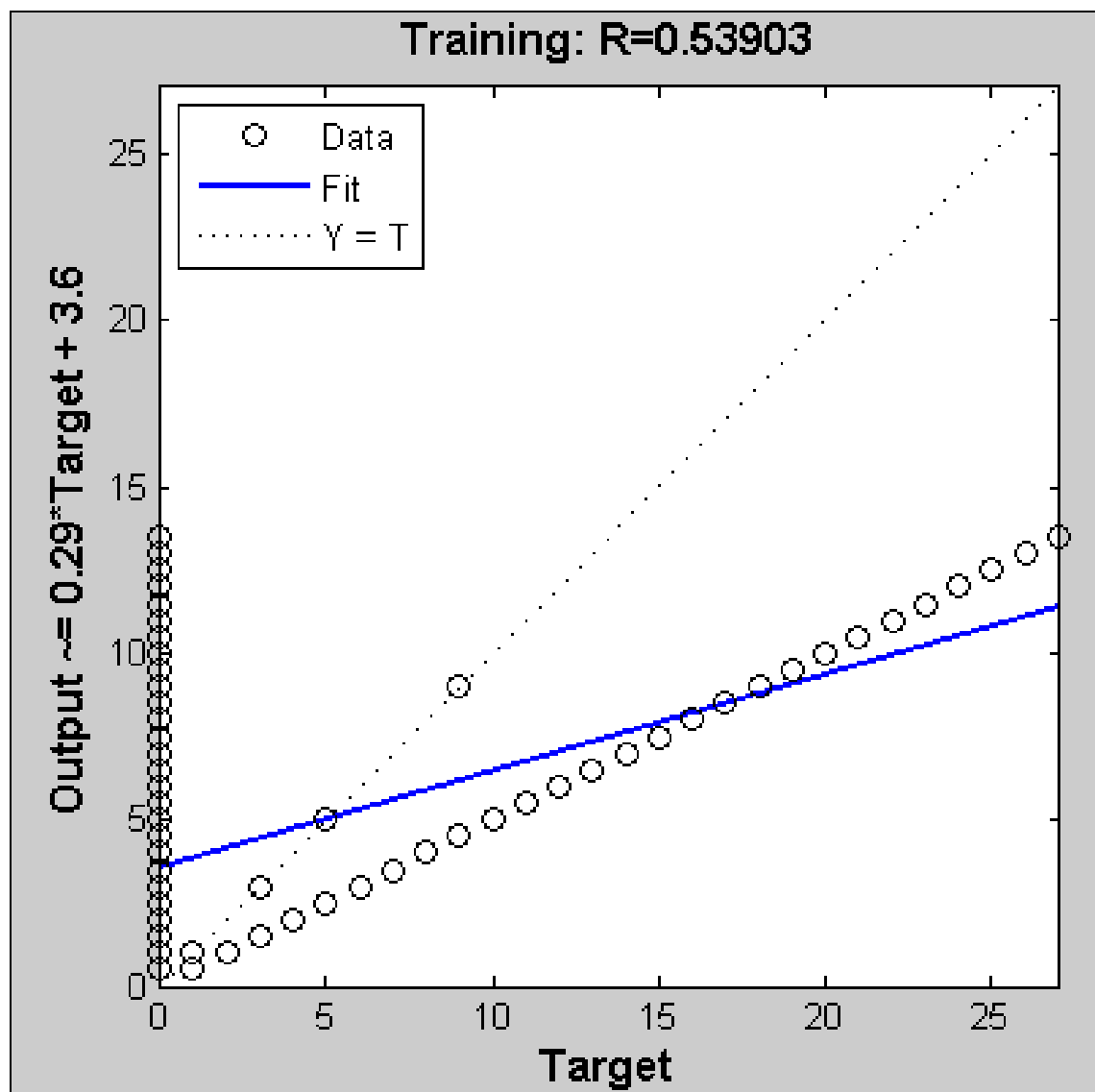


Figure C.12: Trainrp Regression Plot

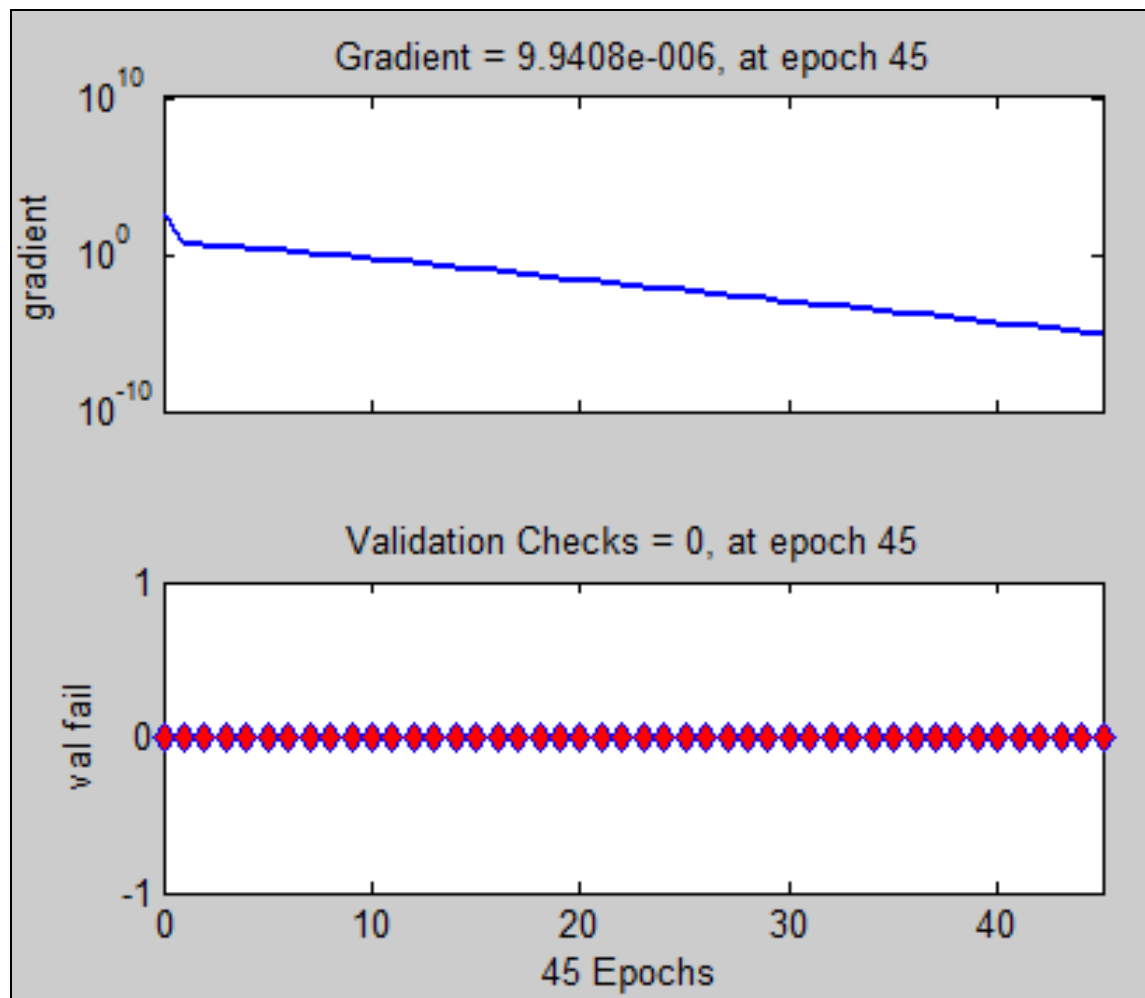


Figure C.13: Trainrp Training Parameters

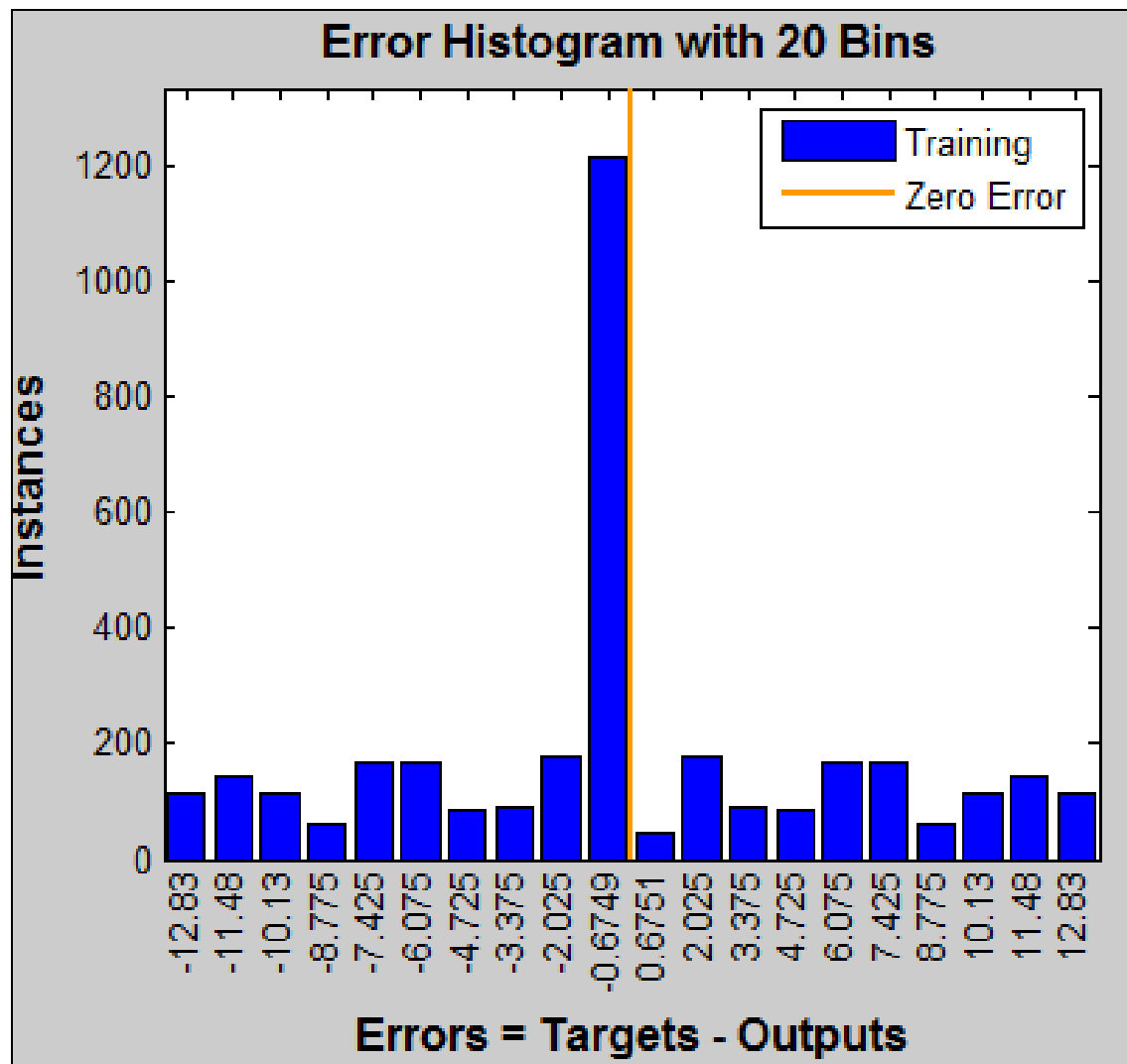


Figure C.14: Trainrp Error Histogram

Appendix D: TVT Graphs

The following figures show the results of the five ANN simulations for each TVT ratio. The 70-15-15 percent and 80-10-10 percent ratios were the only ones to produce linear plots where all of the ANN derived values were positive and the 70-15-15 percent ratio was the only one to produce ANN derived values on the one-to-one slope for all five simulations. The default train training function was used for all of the TVT simulations.

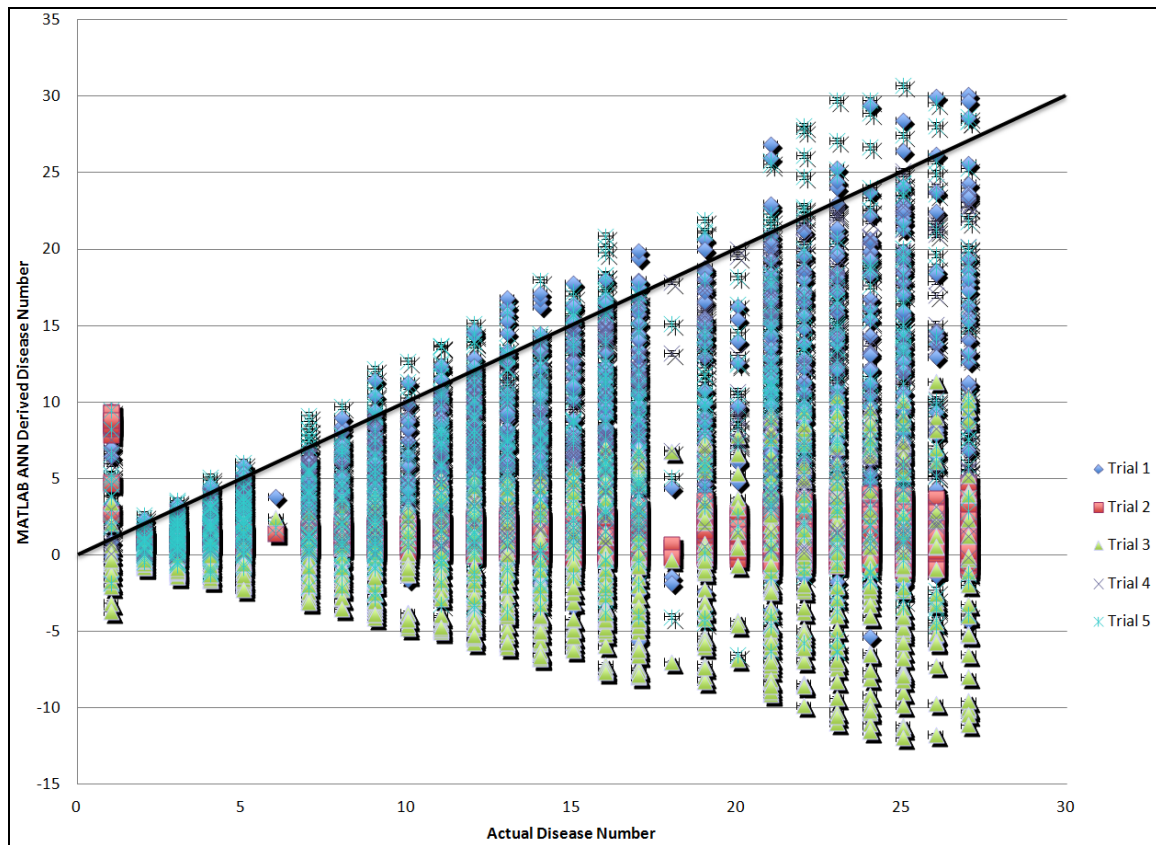


Figure D.1: 50-25-25 % TVT Ratio Plot

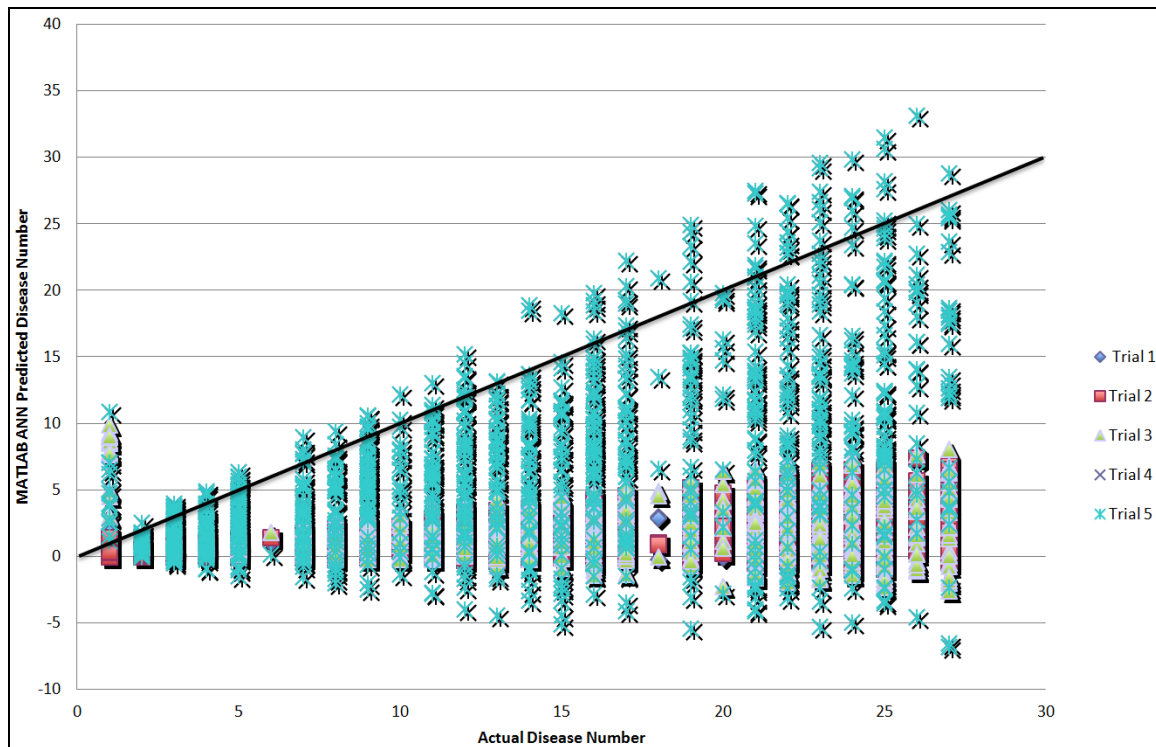


Figure D.2: 60-20-20% TVT Ratio Plot

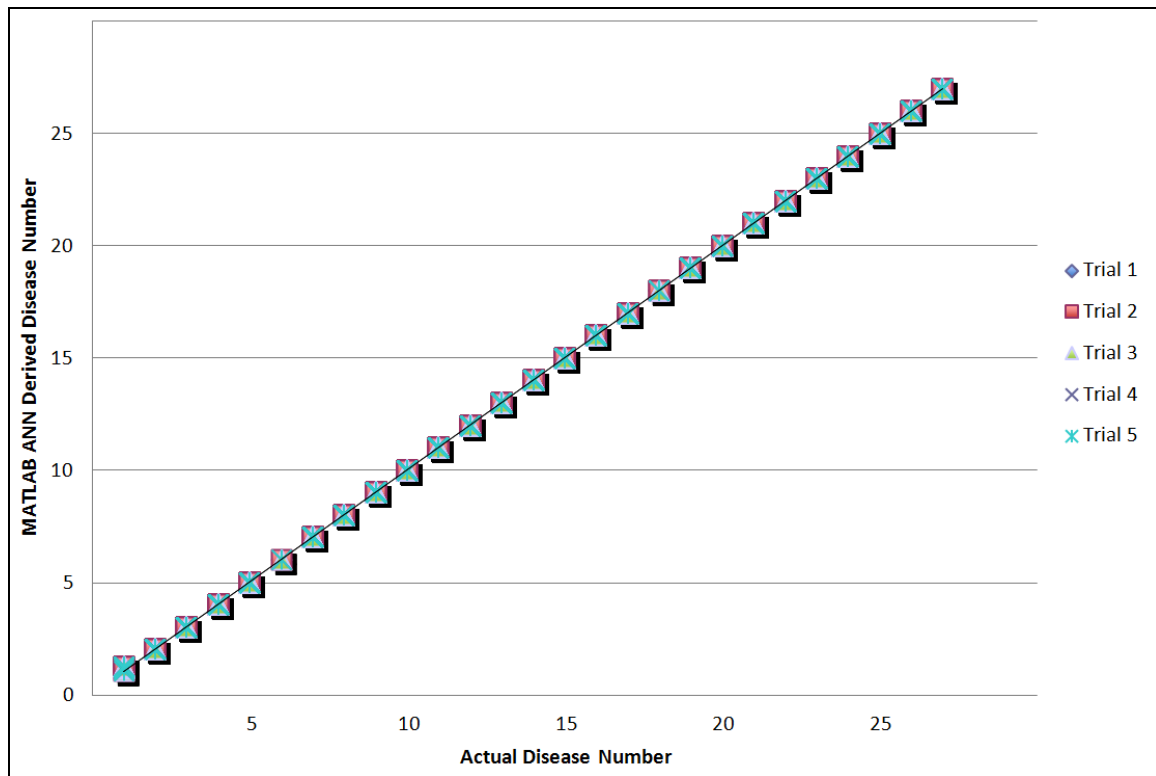


Figure D.3: 70-15-15% TVT Ratio Plot

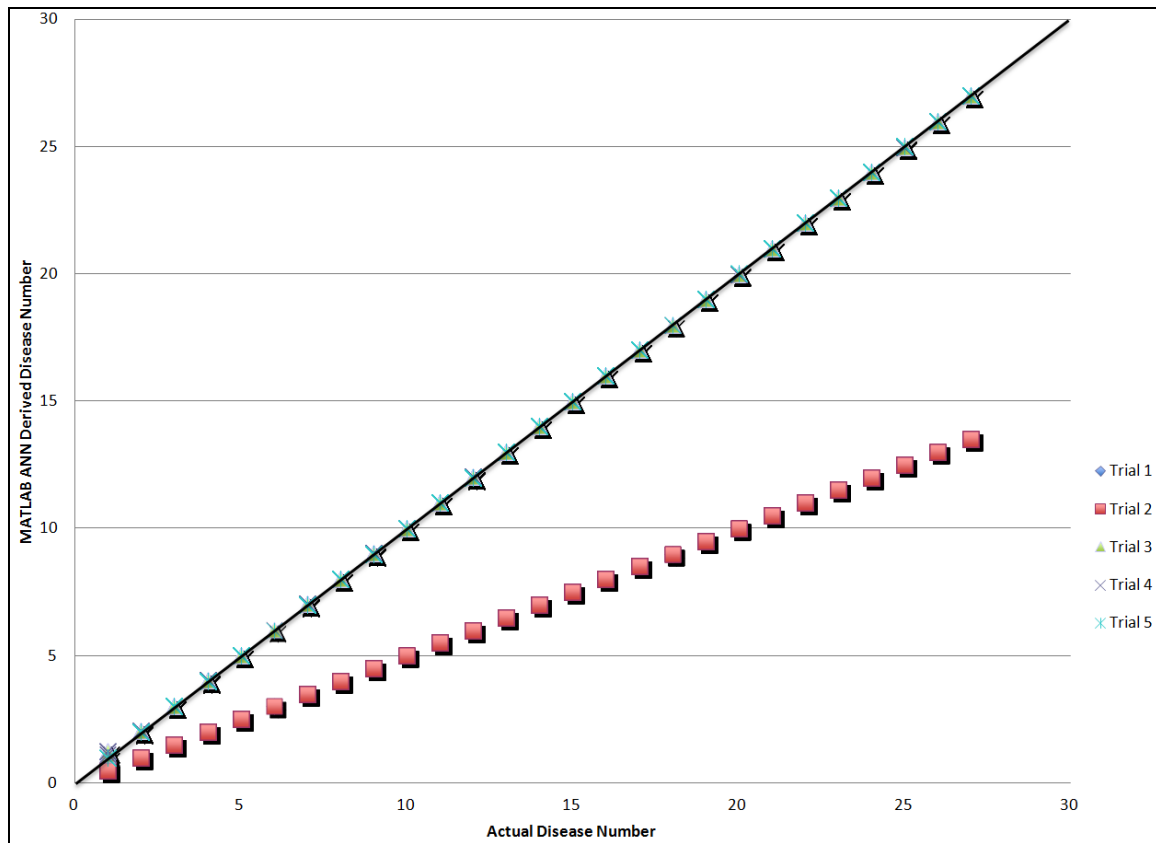


Figure D.4: 80-10-10% TVT Ratio Plot

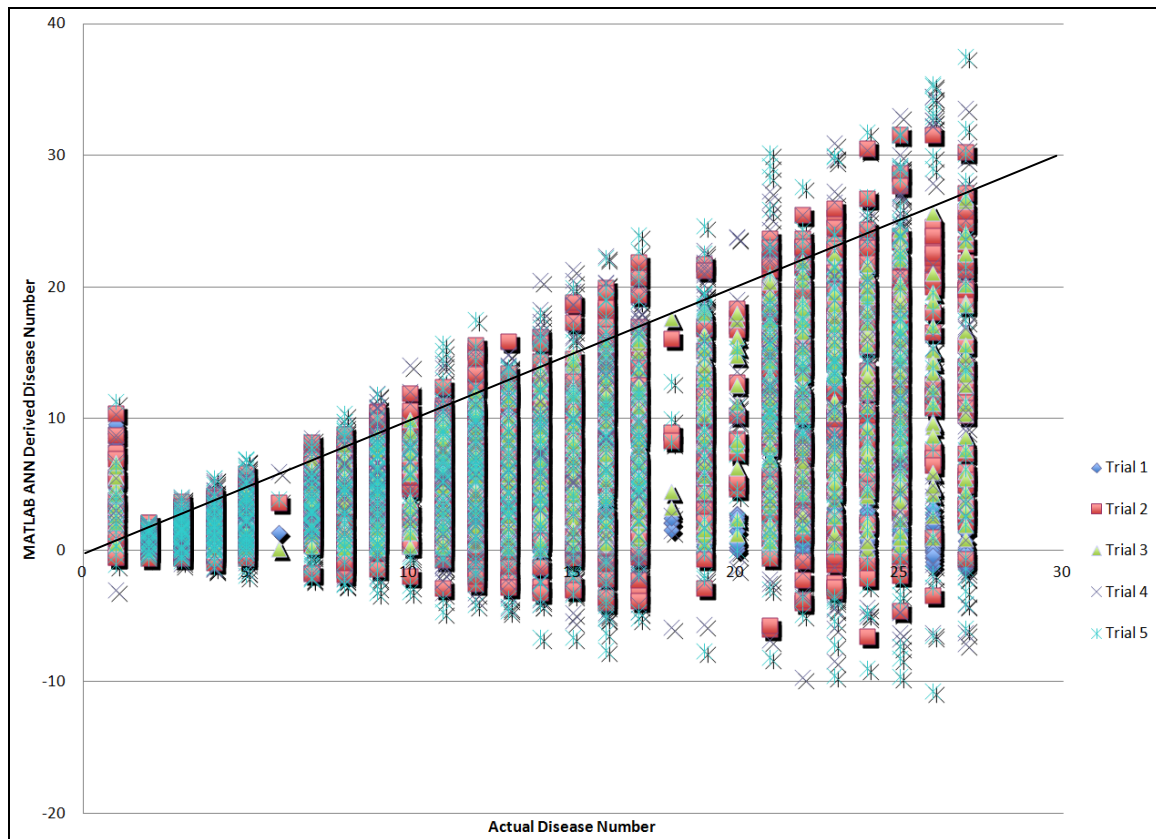


Figure D.5: 90-5-5% TVT Ratio Plot

Appendix E: List of Training Functions

Table E.1 shows each of the training functions used in the ANN simulations along with a brief description of how each function trains the network. All of the descriptions were found in the MATLAB Neural Network Toolbox Reference Guide.

Table E.1: Training Function Descriptions

Training Function	Description
train	Default training function used in the MATLAB ANN. Usually defaults to trainlm functions.
trainb	Trains a network using batch training by updating weight and bias learning rules
trainc	Trains a network cyclical order weight and bias learning rules
traincgb	Trains a network using conjugate gradient backpropagation in conjunction with Powell-Beale restarts to update weight and bias values
traincgf	Trains a network using conjugate gradient backpropagation in conjunction with Fletcher-Reeves updates to update weight and bias values
traincgp	Trains a network using conjugate gradient backpropagation in conjunction with Polak-Ribiere updates to update weight and bias values
traingd	Trains a network using gradient descent backpropagation to update weight and bias values
traingda	Trains a network using gradient descent with adaptive learning rate backpropagation to update weight and bias values
traingdm	Trains a network using gradient descent with momentum backpropagation to update weight and bias values
traingdx	Trains a network using gradient descent with momentum and adaptive learning rate backpropagation to update weight and bias values
trainlm	Trains a network using Levenberg-Marquardt backpropagation to update weight and bias values
trainoss	Trains a network using One-step secant backpropagation to update weight and bias values
trainr	Trains a network using random order incremental training with learning functions to update weight and bias values
trainrp	Trains a network using resilient backpropagation to update weight and bias values
trains	Trains a network using sequential order incremental training with learning functions to update weight and bias values
trainscg	Trains a network using scaled conjugate gradient backpropagation to update weight and bias values

Appendix F: Training Function Plots

The following Figures F.1-F.15 show the actual disease values versus the ANN derived disease values for the 15 training functions tested in the ANN. Each plot shows the three simulations run per training function, as well as a solid black line representing the desired one to one slope. The one to one slope line makes it easier to distinguish whether individual trials and training functions as a whole were able to generate disease values similar to the actual values.

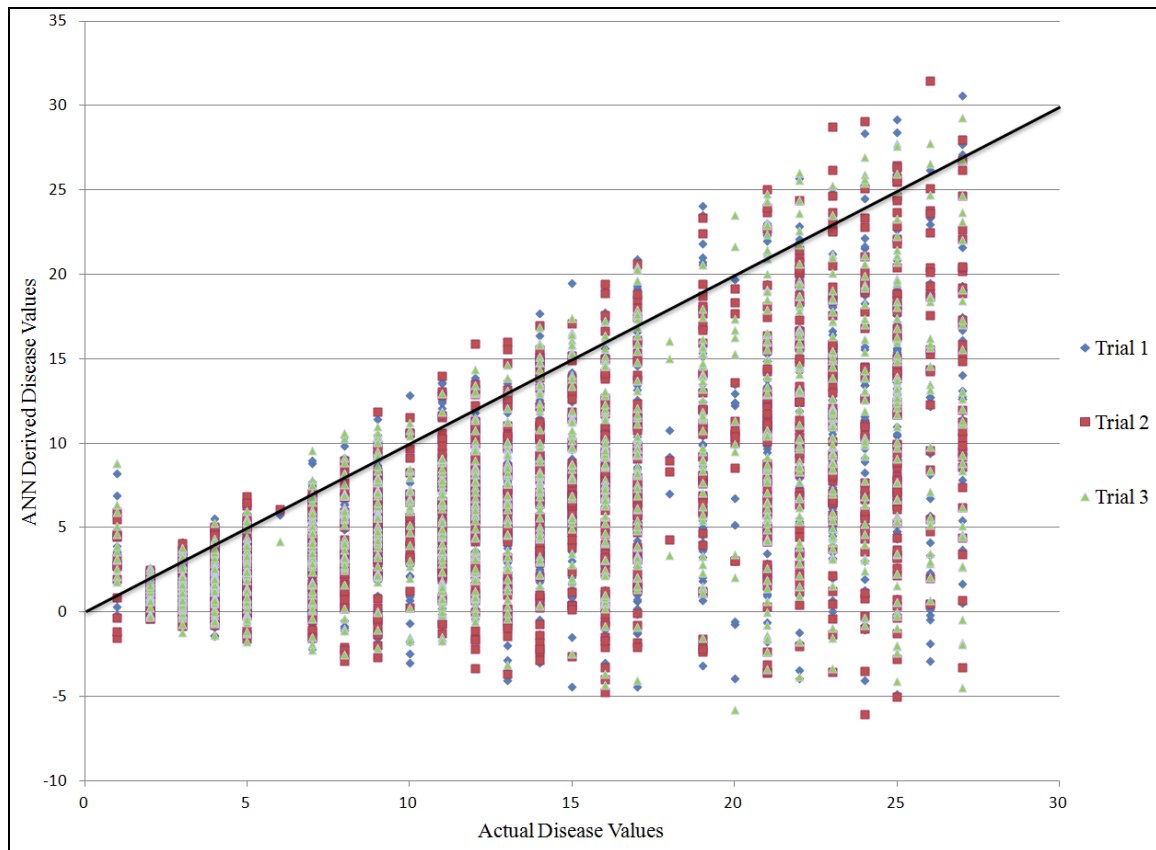


Figure F.1: Trainb Function Disease Plot

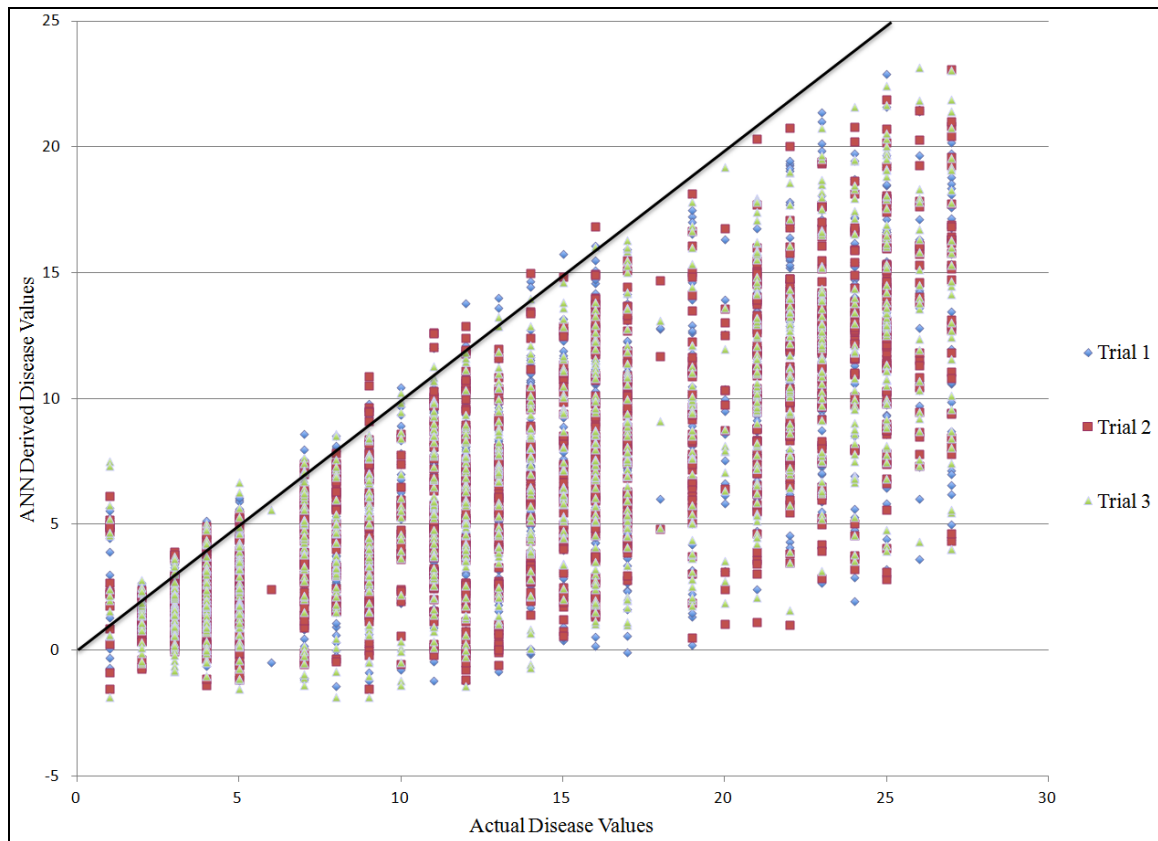


Figure F.2: Trainc Function Disease Plot

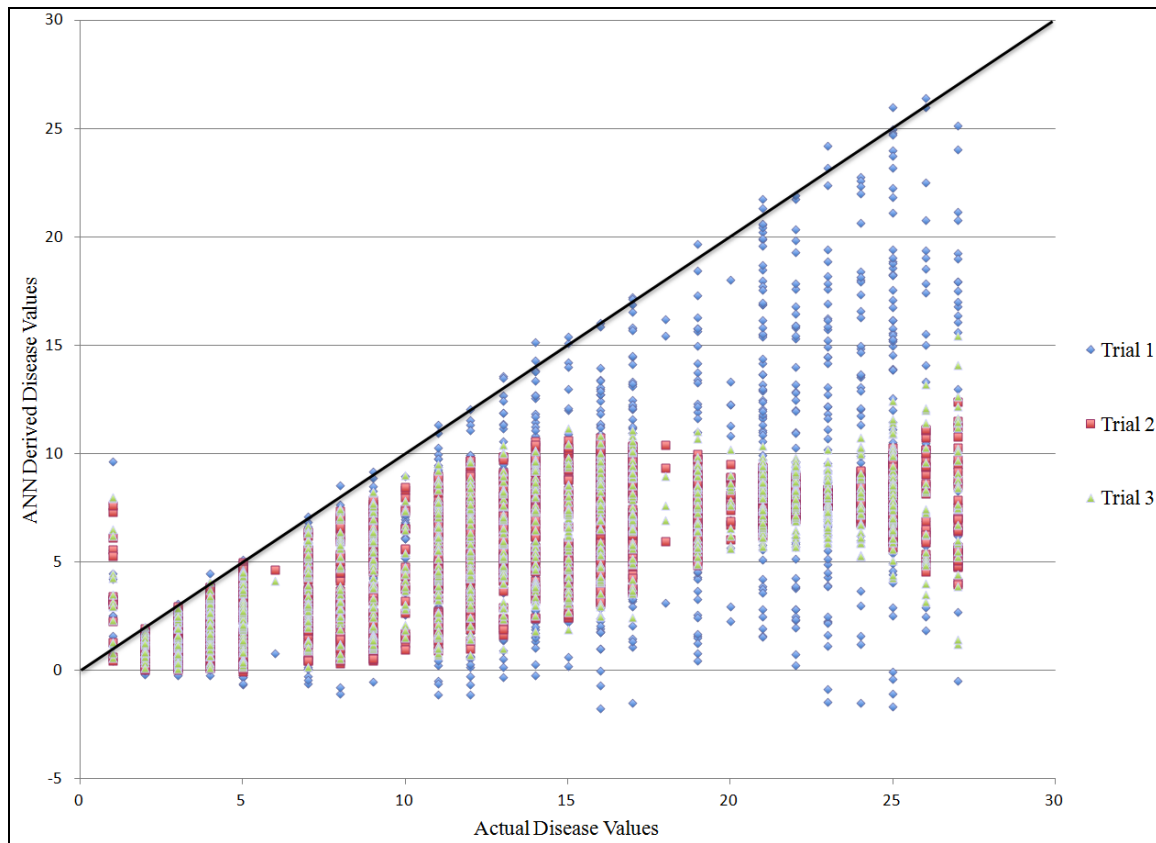


Figure F.3: Traincgb Function Disease Plot

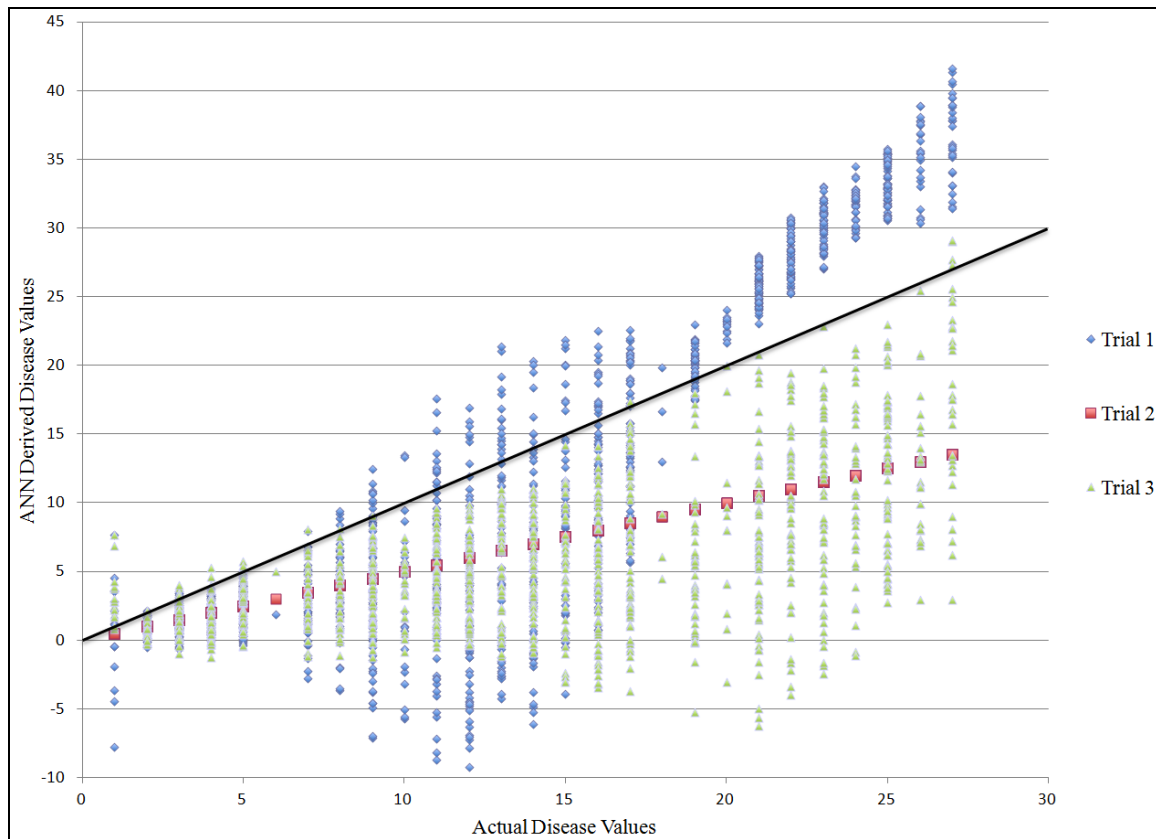


Figure F.4: Traincgrf Function Disease Plot

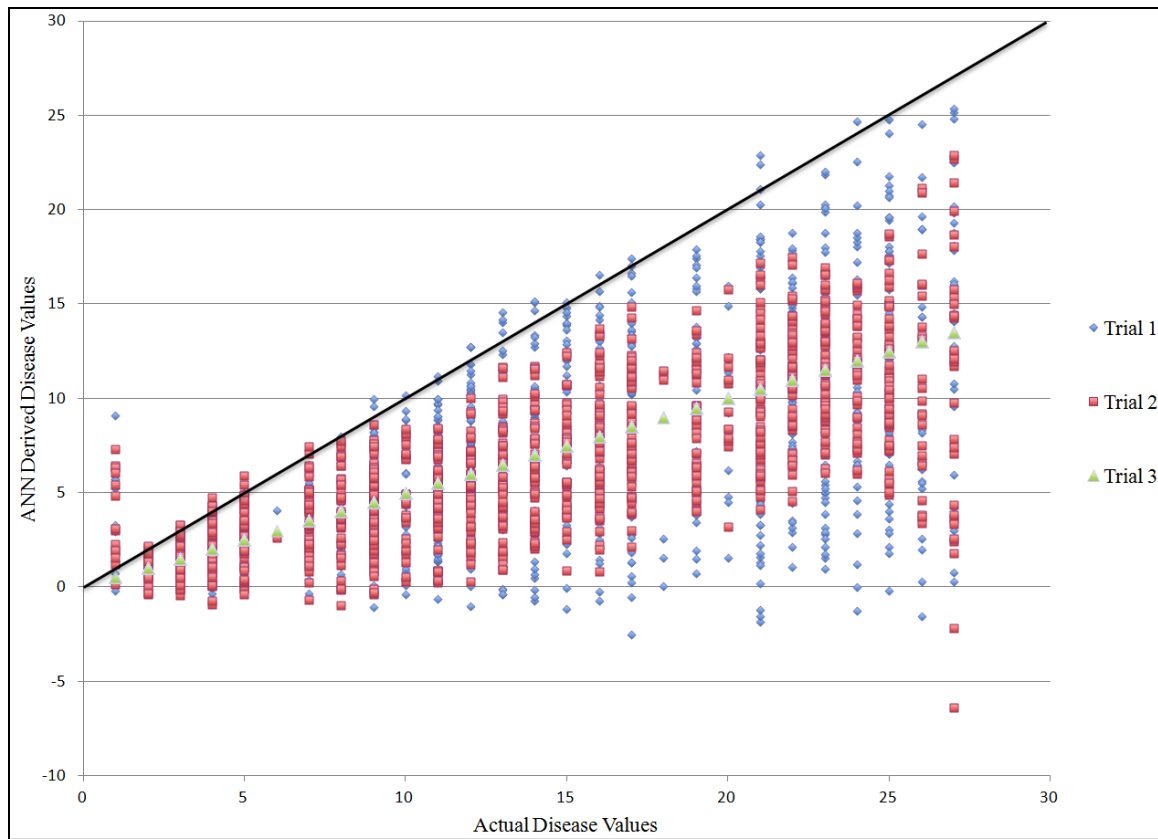


Figure F.5: Traincgp Function Disease Plot

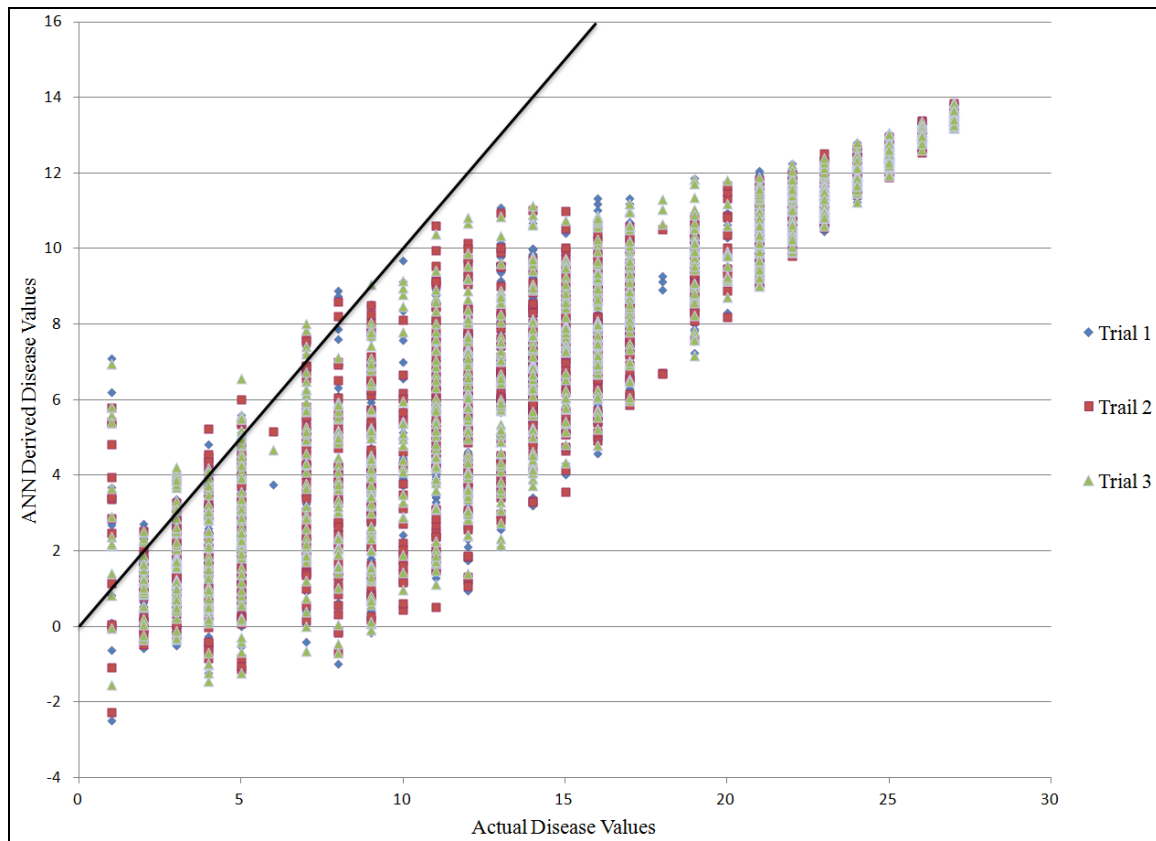


Figure F.6: Traingd Function Disease Plot

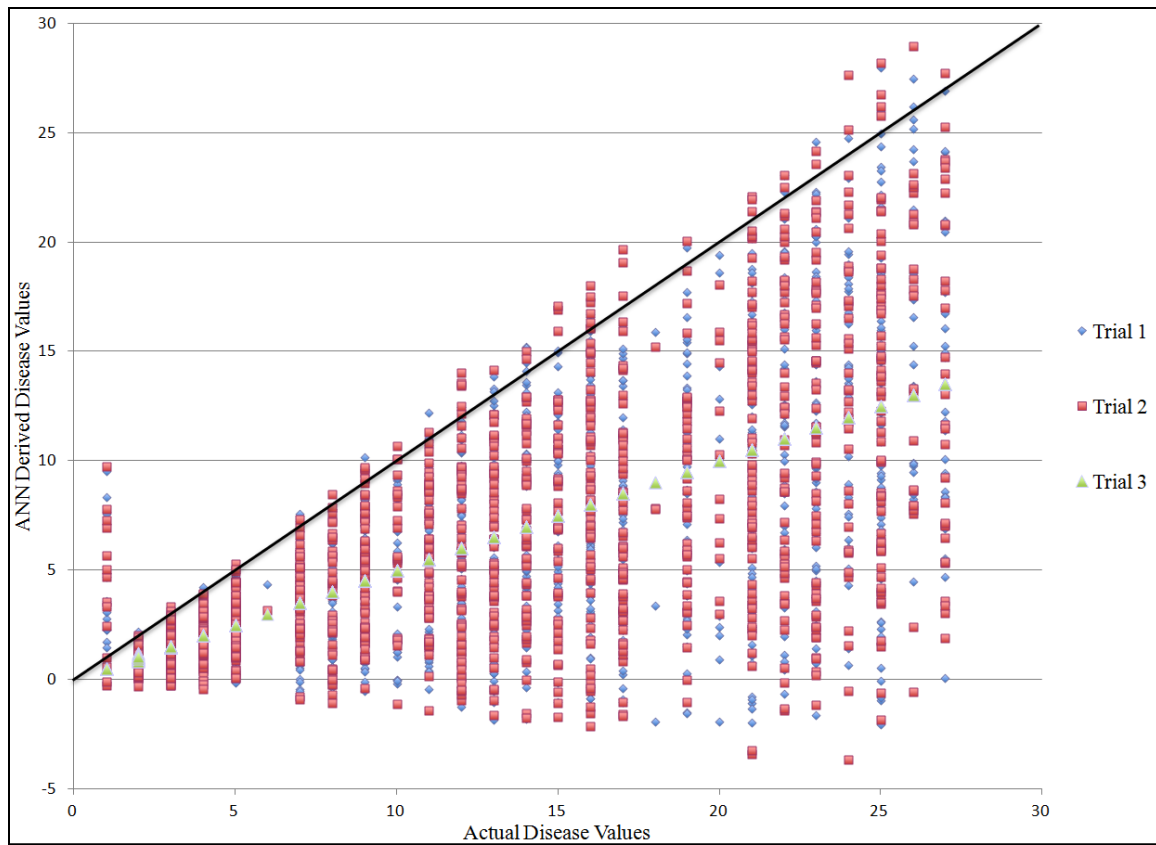


Figure F.7: Traingda Function Disease Plot

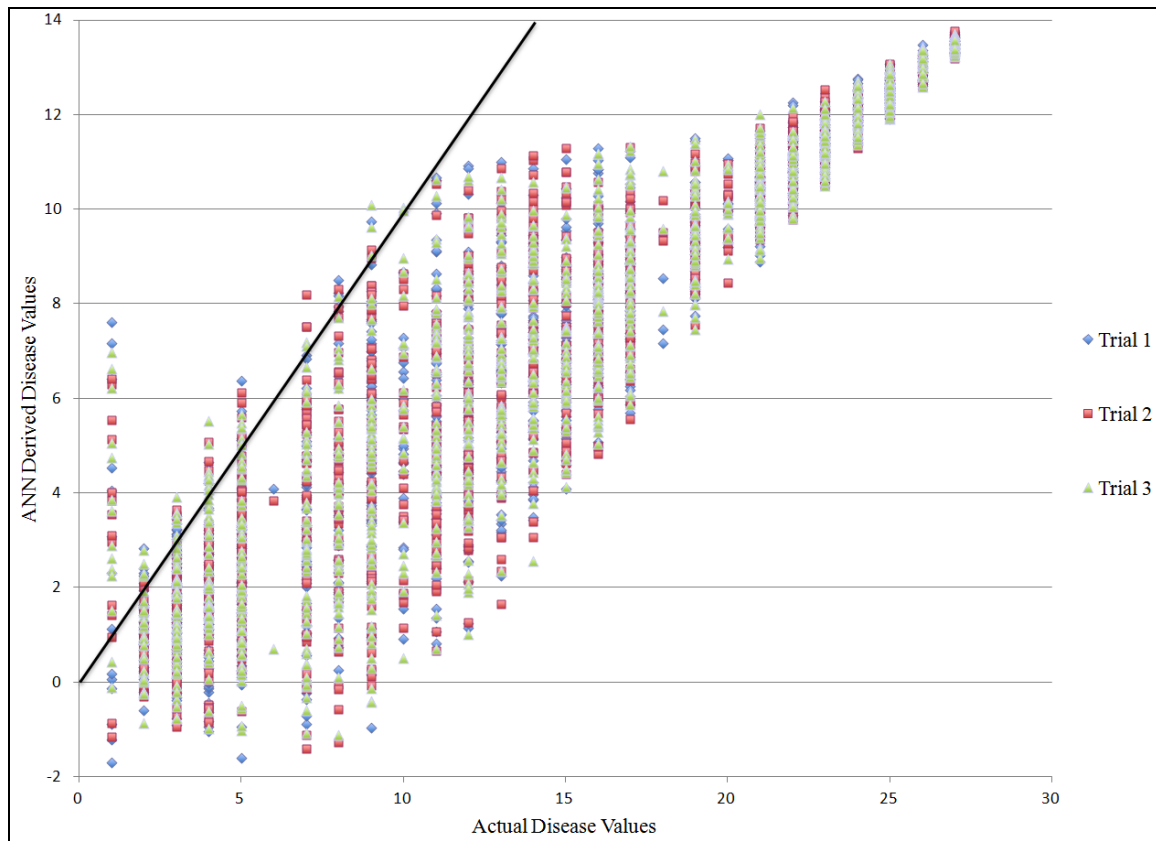


Figure F.8: Traingdm Function Disease Plot

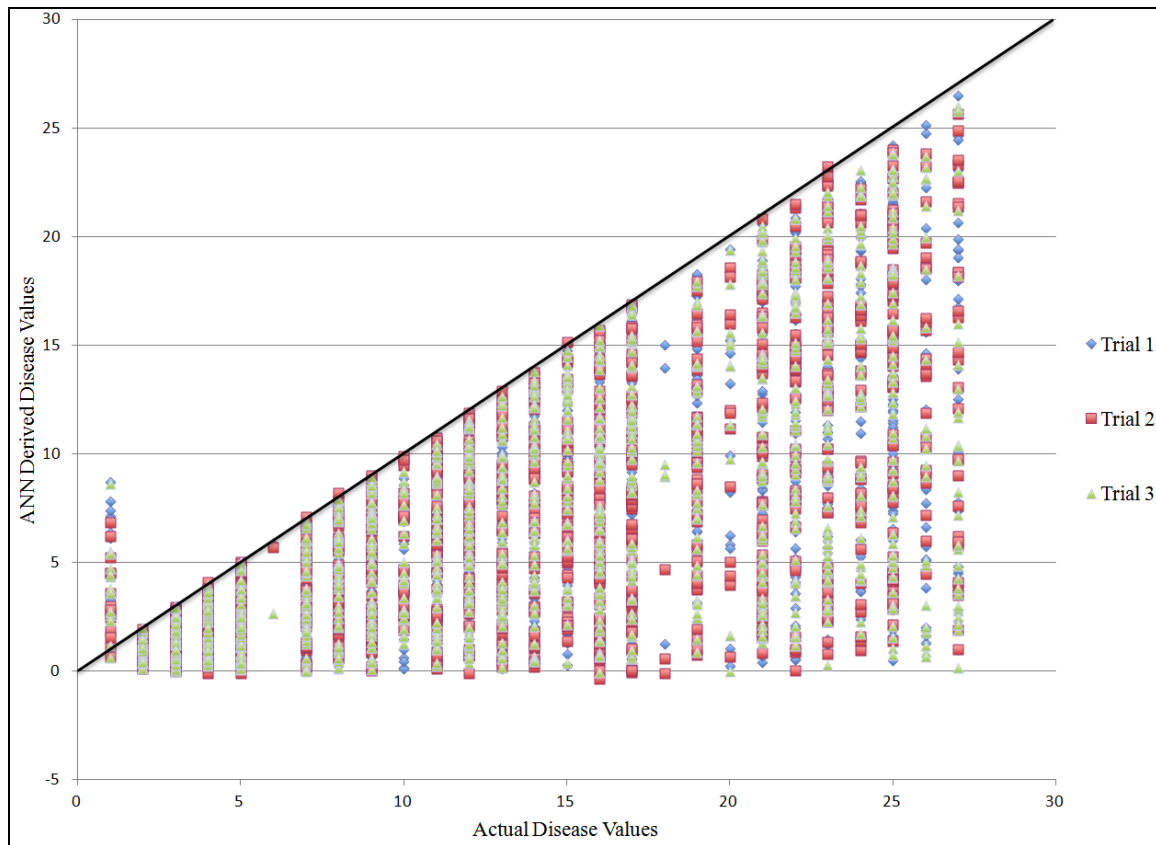


Figure F.9: Traingdx Function Disease Plot

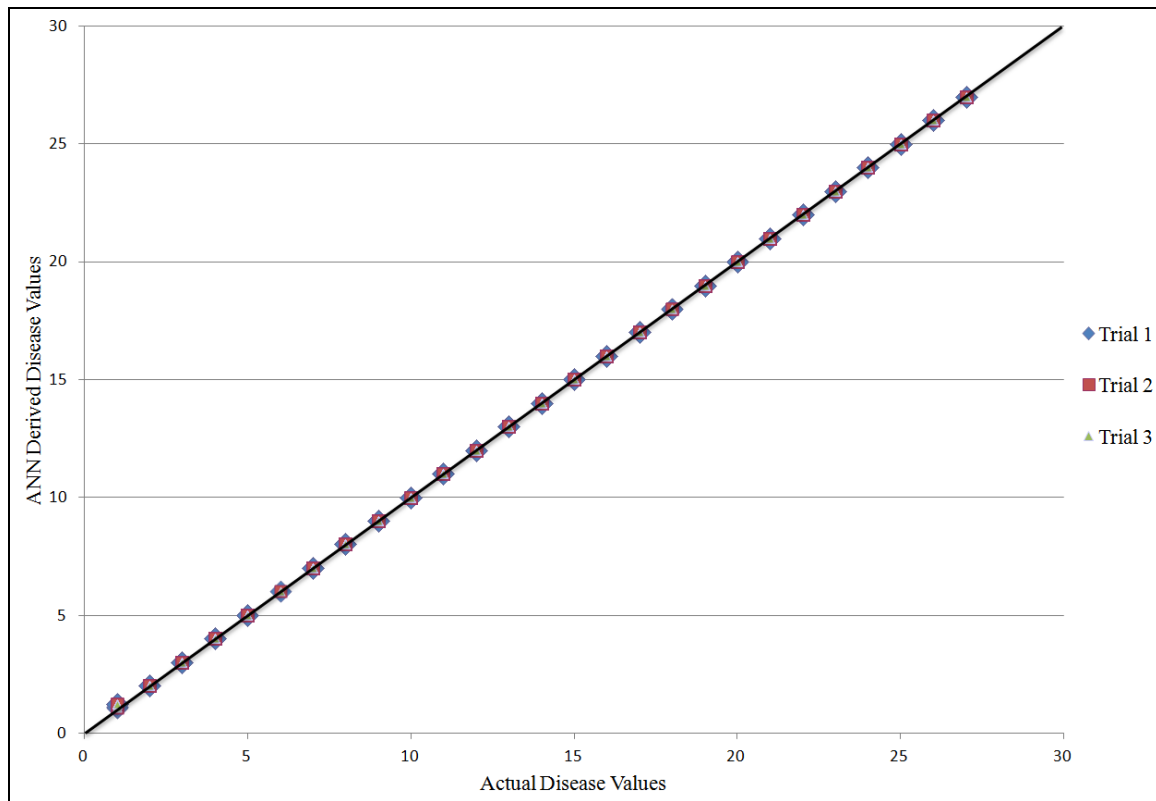


Figure F.10: Trainlm Function Disease Plot

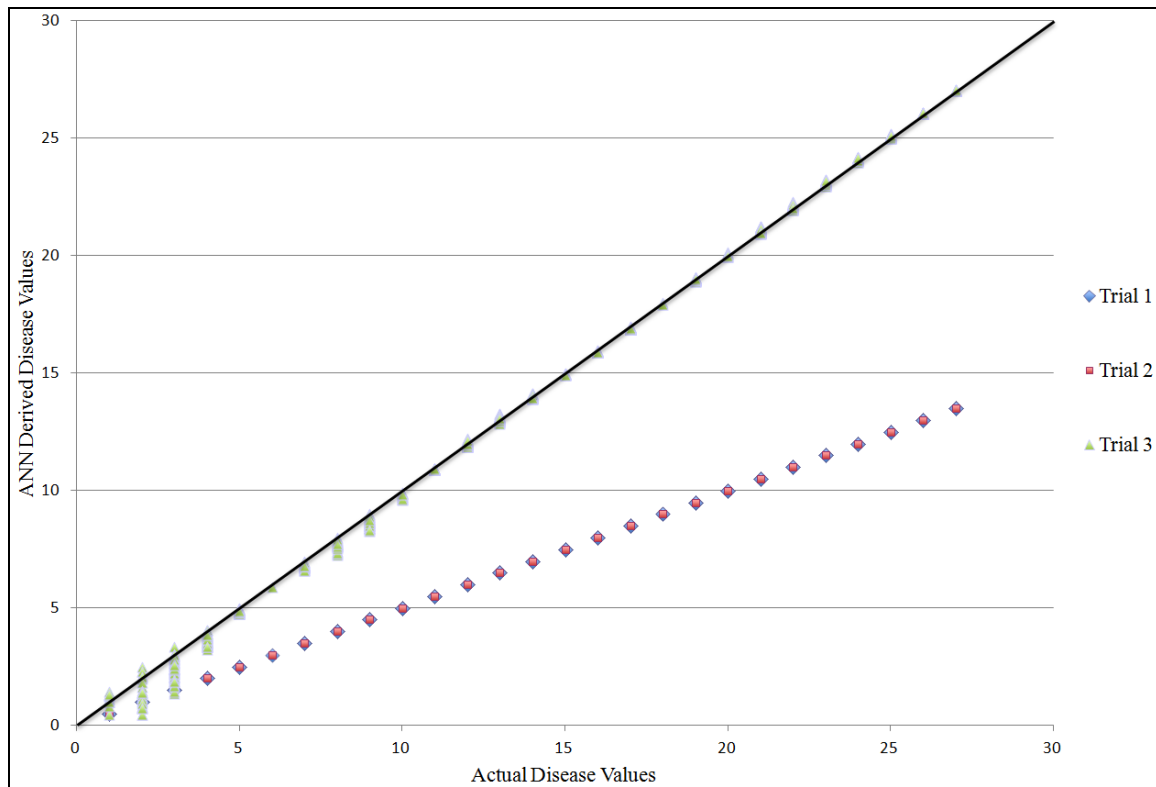


Figure F.11: Trainoss Function Disease Plot

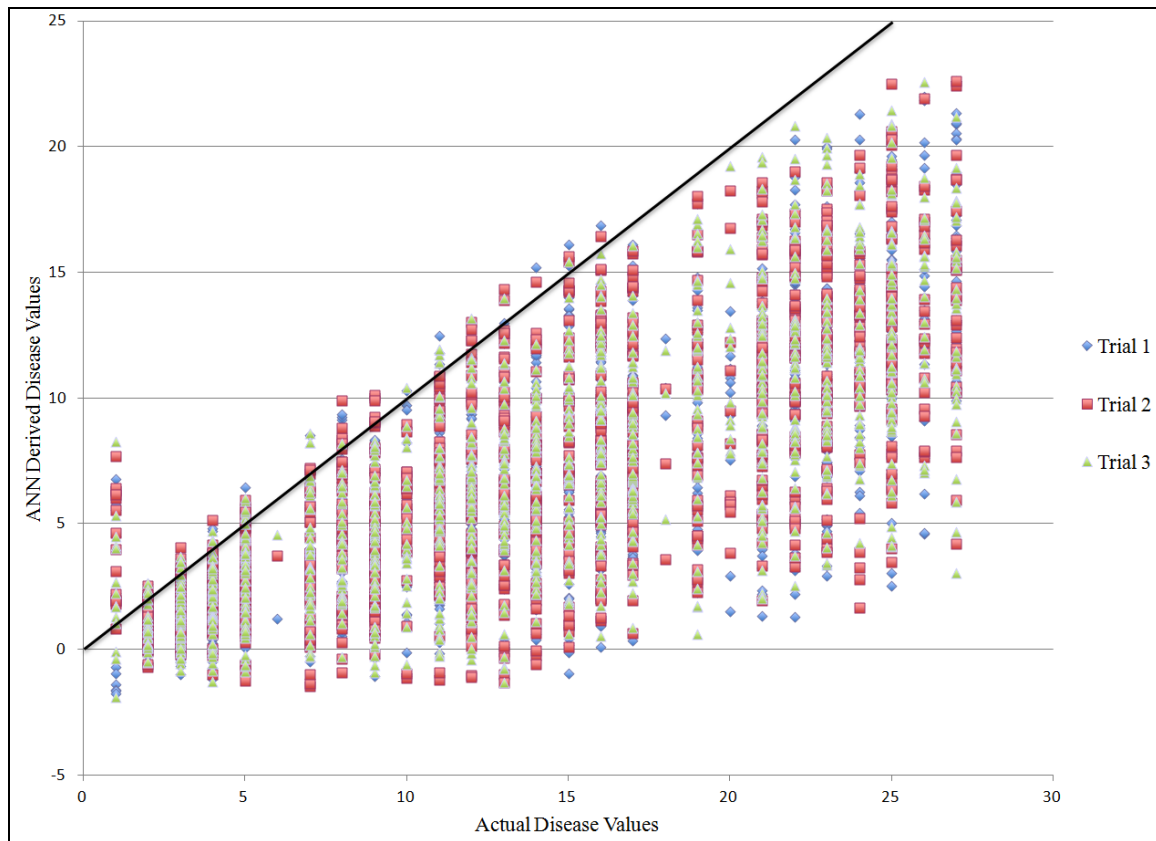


Figure F.12: Trainr Function Disease Plot

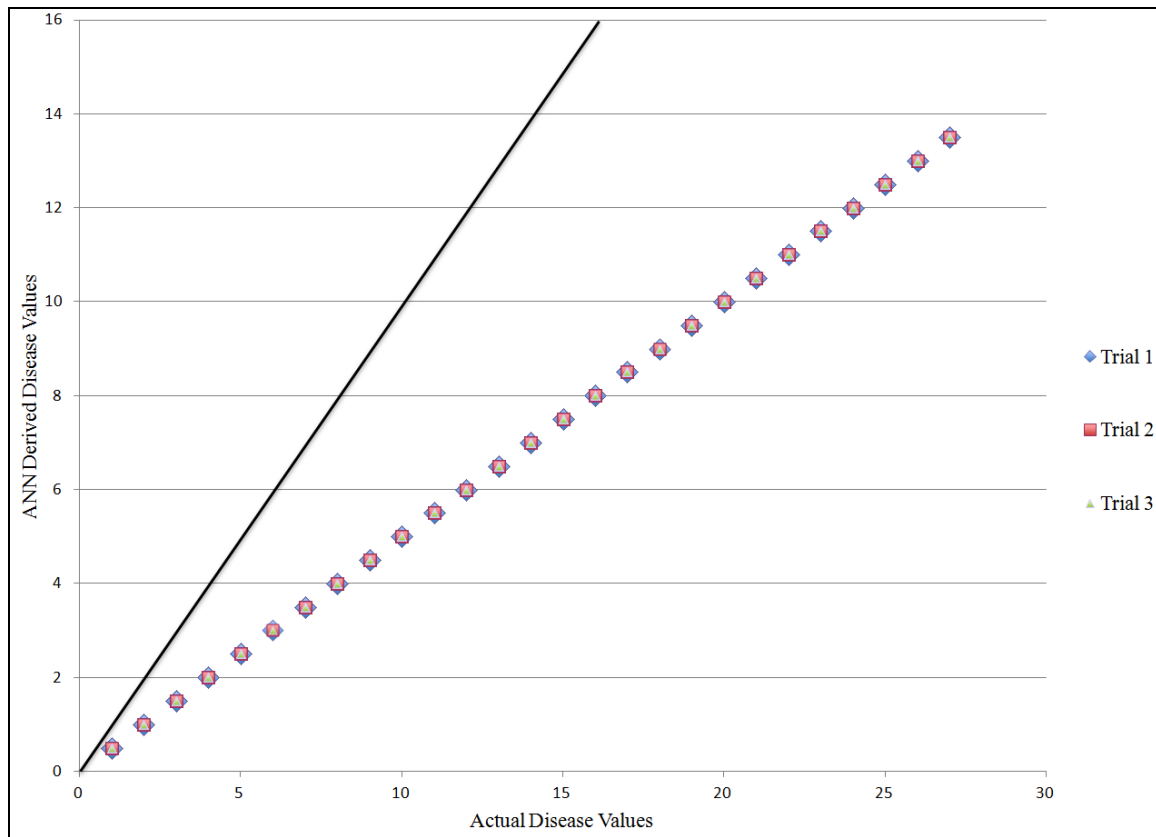


Figure F.13: Trainrp Function Disease Plot

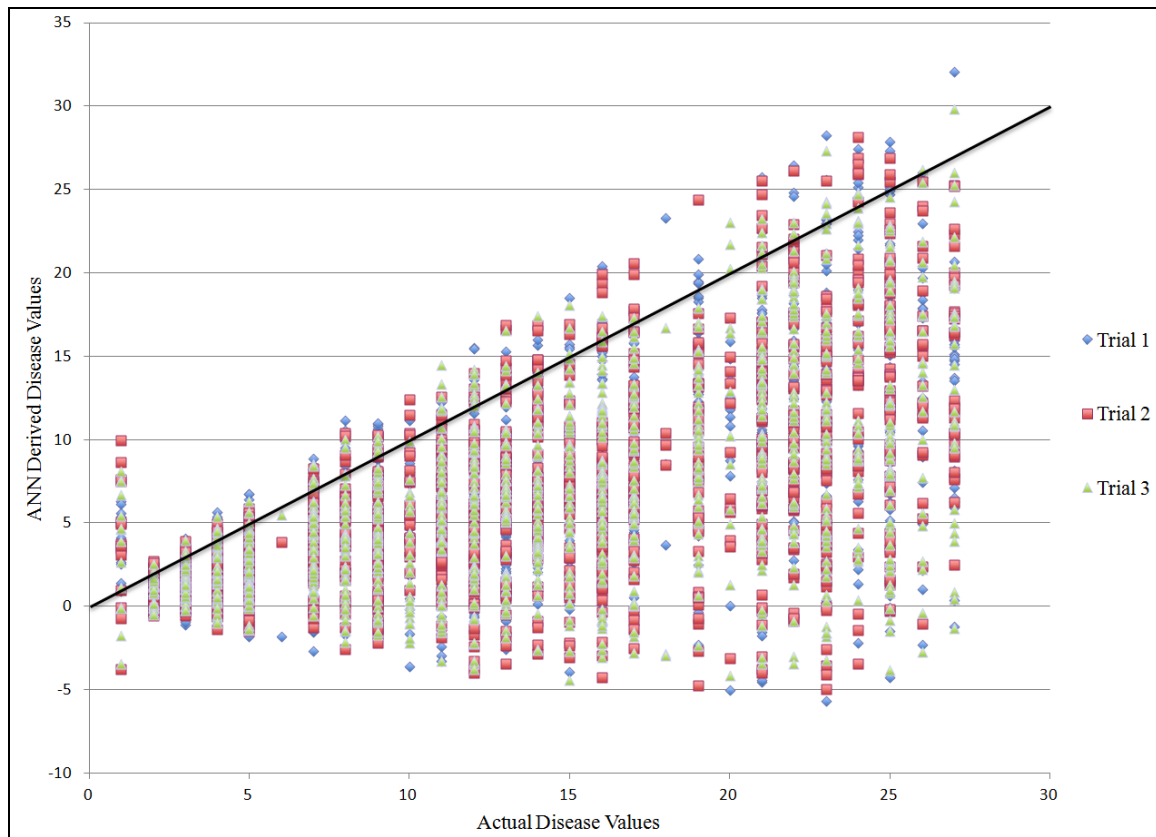


Figure F.14: Trains Function Disease Plot

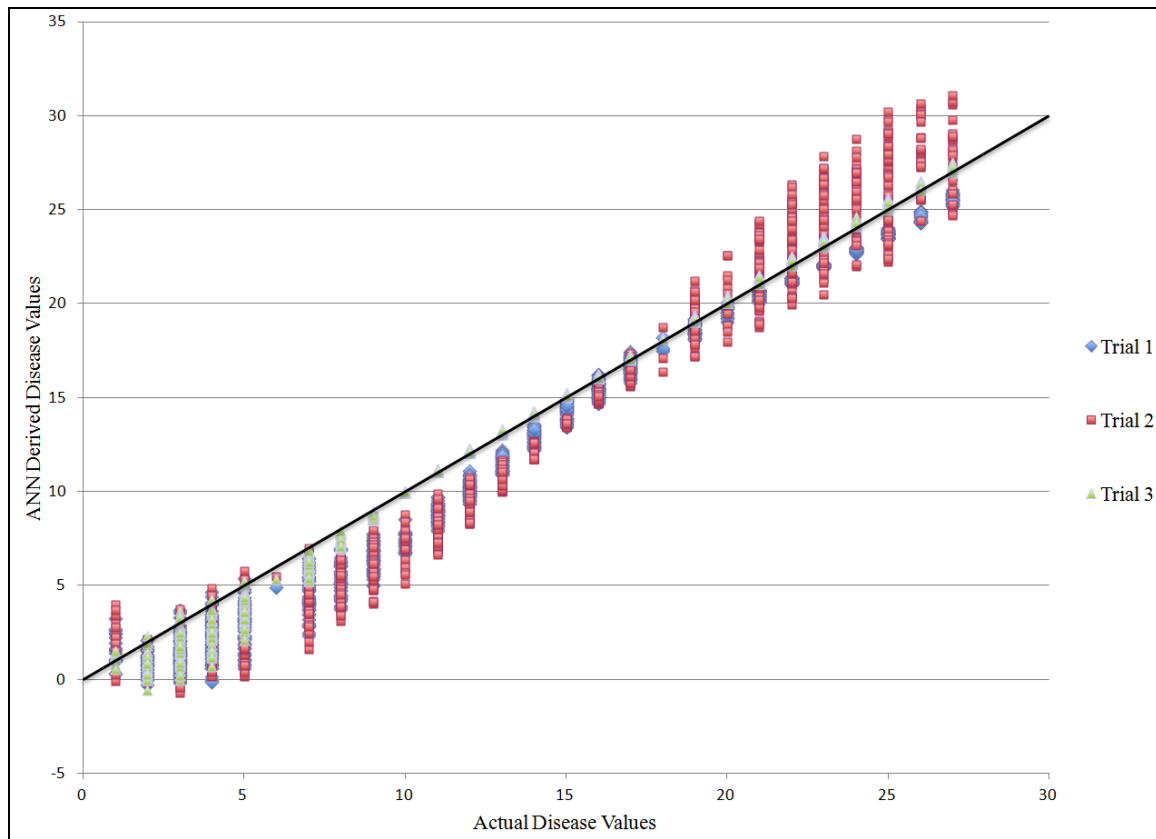


Figure F.15: Trainscg Function Disease Plot

Appendix G: Uncurated chemical data

Table G.1 contains the input data for each of the three uncurated chemicals used to test the predictability of the ANN. The ANN derived species and disease outputs are also shown with the disease outputs round to the nearest whole number. The rounding allowed the predictions to be matched up against the values used for the disease groups. The bolded rounded disease numbers and disease category outputs are those that have literature supporting the networks predictions.

Table G.1: Uncurated Chemical Input, Output, and ANN Derived Disease Number Data

Uncurated Chemical	Molecular Weight	Hydrogen Acceptors	Hydrogen Donors	ANN Derived Species	ANN Derived Disease	Rounded ANN Derived Disease	Disease Category
Cystaphos	179.11	5	2	9.7026	7.1164	7	Endocrine system Diseases
	179.11	5	2	12.9396	9.4451	9	Female Urogenital Diseases/Pregnancy Complications
	179.11	5	2	16.1766	11.7738	12	Male Urogenital Diseases
	179.11	5	2	17.2556	14.7681	15	Cancer - Neoplasms
	179.11	5	2	18.3346	15.2486	15	Cancer - Neoplasms
	179.11	5	2	26.9666	19.5361	20	Parasitic Diseases
	179.11	5	2	4.3076	3.2353	3	Cardiovascular Diseases
	179.11	5	2	5.3866	4.0115	4	Congenital/Hereditary/Neonatal Diseases
	179.11	5	2	9.7026	7.1164	7	Endocrine system Diseases
	179.11	5	2	12.9396	9.4451	9	Female Urogenital Diseases/Pregnancy Complications
6-HO-BDE-47	501.79	2	1	14.0186	10.2214	10	Hemic and Lymphatic Diseases
	501.79	2	1	15.0976	10.9976	11	Immune System Diseases
	501.79	2	1	17.2556	12.5501	13	Mental Disorders
	501.79	2	1	18.3346	13.3263	13	Mental Disorders
	501.79	2	1	20.4926	14.8787	15	Neoplasms (Cancer)
	501.79	2	1	22.6506	16.4312	16	Nervous System Diseases
	501.79	2	1	23.7296	17.2074	17	Nutritional and Metabolic Diseases (2)
	501.79	2	1	26.9666	19.5361	20	Parasitic Diseases
	501.79	2	1	4.0376	3.2353	3	Cardiovascular Diseases
	501.79	2	1	5.3866	4.0115	4	Congenital/Hereditary/Neonatal Diseases
4,4'-diiodobiphenyl	406.00	0	0	9.7026	7.1164	7	Endocrine system Diseases
	406.00	0	0	12.9396	9.4451	9	Female Urogenital Diseases/Pregnancy Complications
	406.00	0	0	15.0976	10.9976	11	Immune System Diseases
	406.00	0	0	17.2556	12.5501	13	Mental Disorders
	406.00	0	0	26.9666	19.5361	20	Parasitic Diseases
	406.00	0	0	1.0037	1.9496	2	Bacterial Infections and Mycoses
	406.00	0	0	3.2286	2.4591	2	Bacterial Infections and Mycoses
	406.00	0	0	7.5446	5.564	6	Disorders of Environmental Origin
	406.00	0	0	8.6236	6.3402	6	Disorders of Environmental Origin
	406.00	0	0	9.7026	7.1164	7	Endocrine system Diseases

References

- Abe, H., Ashizawa, K., Katsuragawa, S., MacMahon, H., & Doi, K. (2002). Use of an artificial neural network to determine the diagnostic value of specific clinical and radiologic parameters in the diagnosis of interstitial lung disease on chest radiographs. *Academic Radiology*, 9(1), 13-17. doi:10.1016/S1076-6332(03)80291-X
- Abe, H., Ashizawa, K., Li, F., Matsuyama, N., Fukushima, A., Shiraishi, J., . . . Doi, K (2004). Artificial neural networks (ANNs) for differential diagnosis of interstitial lung disease: Results of a simulation test with actual clinical cases. *Academic Radiology*, 11(1), 29-37. doi:10.1016/S1076-6332(03)00572-5
- Ahmad, S., & Gromiha, M. (2003). Design and Training of a Neural Network for Predicting the SOLvent Accessibility of Proteins. *Journal of Computational Chemistry* , 1313-1320.
- Ashizawa, K., MacMahon, H., Ishida, T., Nakamura, K., Vyborny, C., Katsuragawa, S., & Doi, K. (1999). Effect of an artificial neural network on radiologists' performance in the differential diagnosis of interstitial lung disease using chest radiographs. *American Journal of Roentgenology*, 172(5), 1311-1315.
- Babaoglu, I., Baykan, O. K., Aygul, N., Ozdemir, K., & Bayrak, M. (2009). Assessment of exercise stress testing with artificial neural network in determining coronary artery disease and predicting lesion localization. *Expert Systems with Applications*, 36(2), 2562-2566. doi:10.1016/j.eswa.2007.11.013
- Barkaia, V. S., Torua, R. A., & Elistratova, Z. H. (1989, September-October). Prevention of Radiation Sickness, Induced by Low-level Ionizing Radiation, by Repeated Injections with Increasing Doses of Chemical Radioprotectors. *Radiobiology* , pp. 638-643.
- Beale, M. H., Hagan, M. T., & Demuth, H. B. (2013). Neural Network Toolbox: User's Guide. Natick, Massachusetts, U.S.A.: The Math Works.
- Biglarian, A., Bakhshi, E., Gohari, M. R., & Khodabakhshi, R. (2012). Artificial neural network for prediction of distant metastasis in colorectal cancer. *Asian Pacific Journal of Cancer Prevention*, 13(3), 927-930. doi:10.7314/APJCP.2012.13.3.927
- Cao, J., Lin, Y., Guo, L. H., Zhang, A. Q., Wei, Y., & Yang, Y. (2010, November 9). Structure-based Investigations on the Binding Interaction of Hydroxylated Polybrominated Diphenyl Ethers with Thyroxine Transport Proteins. *Toxicology* , pp. 1-3.

- Colak, M. C., Colak, C., Kocaturk, H., Sagioglu, S., & Barutcu, I. (2008). Predicting coronary artery disease using different artificial neural network models. *Anadolu Kardiyoloji Dergisi-the Anatolian Journal of Cardiology*, 8(4), 249-254.
- Congressional Digest. (2010, October). Controlling Toxic Substances. *Chemical Safety and Technological Innovation*.
- Cucchetti, A., Vivarelli, M., Heaton, N. D., Phillips, S., Piscaglia, F., Bolondi, L., . . . Pinna, A. D. (2007). Artificial neural network is superior to MELD in predicting mortality of patients with end-stage liver disease. *Gut*, 56(2), 253-258. doi:10.1136/gut.2005.084434
- Dagli, M., & Saritas, I. (2012). Using artificial neural network for the prediction of anemia seen in behcet disease. *Energy Education Science and Technology Part A-Energy Science and Research*, 28(2), 1079-1086.
- Davis AP, Murphy CG, Johnson R, Lay JM, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, King BL, Rosenstein MC, Wiegers TC, Mattingly CJ. The Comparative Toxicogenomics Database: update 2013. *Nucleic Acids Res*. 2013 Jan 1;41(D1):D1104-14.
- Defense Threat Reduction Agency. (2006, October). Experiments on the Biological Action of Neutrons Performed in the Former Soviet Union: A Historical Review. ITT Industries, Inc.
- Deng, X., Li, K., & Liu, S. (1999). Preliminary study on application of artificial neural network to the diagnosis of alzheimer's disease with magnetic resonance imaging. *Chinese Medical Journal*, 112(3), 232-237.
- El-Solh, A., Hsiao, C., Goodnough, S., Serghani, J., & Grant, B. (1999). Predicting active pulmonary tuberculosis using an artificial neural network. *Chest*, 116(4), 968-973. doi:10.1378/chest.116.4.968
- Fang, H., Tong, W., Shi, L., Blair, R., & Perkins, R. (2001). Structure–Activity relationships for a large diverse set of natural, synthetic, and environmental estrogens. *Chemical Research in Toxicology*, 14(3), 280; 280-294; 294.
- Feng, F., Wu, Y., Wu, Y., Nie, G., & Ni, R. (2012). The effect of artificial neural network model combined with six tumor markers in auxiliary diagnosis of lung cancer. *Journal of Medical Systems*, 36(5), 2973-2980. doi:10.1007/s10916-011-9775-1
- Ferrari, S., & Stengel, R. F. (2005). Smooth Function Approximation Using Neural Networks. *IEEE Transactions on Neural Networks*, 24-39.

- Fujikawa, K., Matsui, Y., Kobayashi, T., Miura, K., Oka, H., Fukuzawa, S., . . . Okabe, T. (2003). Predicting disease outcome of non-invasive transitional cell carcinoma of the urinary bladder using an artificial neural network model: Results of patient follow-up for 15 years or longer. *International Journal of Urology*, 10(3), 149-152. doi:10.1046/j.1442-2042.2003.00589.x
- Ghoshal, U. C., & Das, A. (2008). Models for prediction of mortality from cirrhosis with special reference to artificial neural network: A critical review. *Hepatology International*, 2(1), 31-38. doi:10.1007/s12072-007-9026-1
- Gunther, F., & Fritsch, S. (2010). neuralnet: Training of Neural Networks. *The R Journal*, 9.
- Guyon, I. A Scaling Law for the Validation-set Training-set Size Ratio. Berkeley: AT&T Bell Laboratories.
- Hamilton, D., List, A., Butler, T., Hogg, S., & Cawley, M. (2006). Discrimination between parkinsonian syndrome and essential tremor using artificial neural network classification of quantified DaTSCAN data. *Nuclear Medicine Communications*, 27(12), 939-944. doi:10.1097/01.mnm.0000243369.80765.24
- Hendriks, H. S., Antunes Fernandes, E. C., Bergman, A., van den Berg, M., & Westerink, R. H. (2010, December). PCB-47, PBDE-47, and 6-OH-PBDE-47 Differentially Modulate Human GABAA and Alpha4beta2 Nicotinic Acetylcholine Receptors . *Toxicological Sciences* .
- Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (2012). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 64, Supplement(0), 4-17. doi:10.1016/j.addr.2012.09.019
- Lux, A., Mueller, R., Tulk, M., Olivieri, C., Zarrabeita, R., Salonikios, T., & Wirtzner, B. (2013). HHT diagnosis by mid-infrared spectroscopy and artificial neural network analysis. *Orphanet Journal of Rare Diseases*, 8, 94. doi:10.1186/1750-1172-8-94
- Matake, K., Yoshimitsu, K., Kumazawa, S., Higashida, Y., Irie, H., Asayama, Y., . . . Honda, H. (2006). Usefulness of artificial neural network for differential diagnosis of hepatic masses on CT images. *Academic Radiology*, 13(8), 951-962. doi:10.1016/j.acra.2006.04.009
- Matsui, Y., Egawa, S., Tsukayama, C., Terai, A., Kuwao, S., Baba, S., & Arai, Y. (2002). Artificial neural network analysis for predicting pathological stage of clinically localized prostate cancer in the japanese population. *Japanese Journal of Clinical Oncology*, 32(12), 530-535. doi:10.1093/jjco/hyf114

- Matsuki, Y., Nakamura, K., Watanabe, H., & Aoki, T. (2002). Usefulness of an artificial neural network for differentiating benign from malignant pulmonary nodules on high-resolution CT: Evaluation with receiver operating characteristic analysis. *American Journal of Roentgenology*, 178(3), 657-663.
- National Science Foundation. (2013). EPA/NSF Networks for Characterizing Chemical Lifecycle. Arlington, Virginia, U.S.A.
- Nguyen, T., Malley, R., Inkelis, S., & Kuppermann, N. (2002). Comparison of prediction models for adverse outcome in pediatric meningococcal disease using artificial neural network and logistic regression analyses. *Journal of Clinical Epidemiology*, 55(7), 687-695. doi:10.1016/S0895-4356(02)00394-3
- Ning, G., Su, J., Li, Y., Wang, X., Li, C., Yan, W., & Zheng, X. (2006). Artificial neural network based model for cardiovascular risk stratification in hypertension. *Medical & Biological Engineering & Computing*, 44(3), 202-208. doi:10.1007/s11517-006-0028-2
- Orunesu, E., Bagnasco, M., Salmaso, C., Altrinetti, V., Bernasconi, D., Del Monte, P., . . . Mela, G. (2004). Use of an artificial neural network to predict graves' disease outcome within 2 years of drug withdrawal. *European Journal of Clinical Investigation*, 34(3), 210-217. doi:10.1111/j.1365-2362.2004.01318.x
- Raoufy, M. R., Eftekhari, P., Gharibzadeh, S., & Masjedi, M. R. (2011). Predicting arterial blood gas values from venous samples in patients with acute exacerbation chronic obstructive pulmonary disease using artificial neural network. *Journal of Medical Systems*, 35(4), 483-488. doi:10.1007/s10916-009-9384-4
- Ren, S. (2002). Classifying class I and class II compounds by hydrophobicity and hydrogen bonding descriptors. *Environmental Toxicology*, 17(5), 415-423. doi:10.1002/tox.10074
- Salvi, M., Dazzi, D., Pellistri, I., Neri, F., & Wall, J. (2002). Classification and prediction of the progression of thyroid-associated ophthalmopathy by an artificial neural network. *Ophthalmology*, 109(9), 1703-1708. doi:10.1016/S0161-6420(02)01127-2
- Santos-Garcia, G., Varela, G., Novoa, N., & Jimenez, M. (2004). Prediction of postoperative morbidity after lung resection using an artificial neural network ensemble. *Artificial Intelligence in Medicine*, 30(1), 61-69. doi:10.1016/S0933-3657(03)00059-9
- Schultz, T. W. (2002). Structure-activity relationships for gene activation oestrogenicity: Evaluation of a diverse set of aromatic chemicals. *Environmental Toxicology*, 17(1), 14; 14-23; 23

- Seguritan, V., Alves, N., Arnoult, M., Raymond, A., Lorimer, D., Burgin, A. B., et al. (2012). Artificial Neural Networks Trained to Detect Viral and Phage Structural Proteins. *PLOS Computational Biology*.
- Sheppard, D., McPhee, D., Darke, C., Shrethra, B., Moore, R., Jurewitz, A., & Gray, A. (1999). Predicting cytomegalovirus disease after renal transplantation: An artificial neural network approach. *International Journal of Medical Informatics*, 54(1), 55-76. doi:10.1016/S1386-5056(98)00169-5
- Singh, D. V., Maheshwari, G., Shrivastav, R., & Mishra, D. K. (2011). Neural Network - Comparing the Performances of the Training Functions for Predicting the Value of Specific Heat of Refrigerant in Vapor Absorption Refrigeration System. *International Journal of Computational applications*, 5
- Song, J., Venkatesh, S., Conant, E., Arger, P., & Sehgal, C. (2005). Comparative analysis of logistic regression and artificial neural network for computer-aided diagnosis of breast masses. *Academic Radiology*, 12(4), 487-495. doi:10.1016/j.acra.2004.12.016
- Stephan, C., Kahrs, A., Cammann, H., Lein, M., Schrader, M., Deger, S., . . . Jung, K. (2009). A [-2]proPSA-based artificial neural network significantly improves differentiation between prostate cancer and benign prostatic diseases. *Prostate*, 69(2), 198-207. doi:10.1002/pros.20872
- Svetnik, V., Liaw, A., Tong, C., Culberson, J., Sheridan, R., & Feuston, B. (2003). Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, 43(6), 1947-1958. doi:10.1021/ci034160g
- United States Environmental Protection Agency. (2013, February 25). *Basic Information*. Retrieved December 12, 2013, from TSCA Chemical Substance Inventory: <http://www.epa.gov/oppt/existingchemicals/pubs/tscainventory/basic.html#howare>
- Viazzi, F., Leoncini, G., Sacchi, G., Parodi, D., Ratto, E., Falqui, V., . . . Pontremoli, R. (2006). Predicting cardiovascular risk using creatinine clearance and an artificial neural network in primary hypertension. *Journal of Hypertension*, 24(7), 1281-1286. doi:10.1097/01.hjh.0000234107.08368.e5
- Wu, T., He, M., Zang, X., Zhou, Y., Qiu, T., Pan, S., & Xu, X. (2013). A structure-activity relationship study of flavonoids as inhibitors of E. coli by membrane interaction effect. *Biochimica Et Biophysica Acta-Biomembranes*, 1828(11), 2751-2756. doi:10.1016/j.bbamem.2013.07.029

Yamada-Okabe, T., Sakai, H., Kashima, Y., & Yamada-Okabe, H. (2005, January 15). Modulation at a Cellular Level of the Thyroid Hormone Receptor-mediated Gene Expression by 1,2,5,6,9,10-hexabromocyclododecane (HBCD), 4,4'-diiodobiphenyl (DIB), and Nitrofen (NIP). *Toxicology Letters* , pp. 127-133.

Vita

Captain Edward Brouch was born in Aurora, IL and graduated from West Aurora High School. He attended Valparaiso University in Valparaiso, IN to study civil engineering while also joining the Air Force Reserve Officer Training Corps through Detachment 225 at Notre Dame University. After graduating and commissioning in May 2009, Capt Brouch was assigned to the 30th Civil Engineer Squadron at Vandenberg AFB, CA where he served as the a project manager and the Officer in Charge of Mission Engineering Element. He entered the Graduate School of Engineering and Management at the Air Force Institute of Technology in September 2012. After graduation, Capt Brouch will be assigned to the 2nd Civil Engineer Squadron at Barksdale AFB, LA.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 074-0188	
<p>The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of the collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.</p> <p>PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.</p>					
1. REPORT DATE (DD-MM-YYYY) Grad Date: 27-03-2014		2. REPORT TYPE Master's Thesis		3. DATES COVERED (From – To) 01 Oct 2012 – 27 Mar 2014	
4. TITLE AND SUBTITLE Artificial Neural Network Prediction of Chemical-Disease Relationships using Readily Available Chemical Properties				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Brouch, Edward, J., Captain, USAF				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S) Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/ENV) 2950 Hobson Way, Building 640 WPAFB OH 45433-8865				8. PERFORMING ORGANIZATION REPORT NUMBER AFIT-ENV-14-M-12	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Intentionally Left Blank				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Distribution Statement A. Approved for Public Release; Distribution Unlimited Other requests for this document shall be referred to AFIT ENV Office 2950 Hobson Way Wright-Patterson AFB, OH 45433-7765					
13. SUPPLEMENTARY NOTES This material is declared a work of the United States Government and is not subject to copyright protections in the United States					
14. ABSTRACT <p>The natural environment is burdened with a broad range of toxic chemicals, and there is a need to develop a tool that can accelerate the pace at which we learn how chemicals impact disease. This work developed an artificial neural network (ANN) based model that constructed chemical-disease relationships for chemicals found in the Comparative Toxicogenomics Database. A new chemical classification system, based on the molecular weight, hydrogen donors, and hydrogen acceptors, was created to identify chemicals with a unique number that is directly related to these structural properties of the chemical. Diseases were grouped into 27 categories and the chemical-disease associations were made between the chemical and its associated disease category. The ANN model was successfully trained and tested to associated 75 chemical with the 27 disease categories. Simulations with training-validation-testing ratios of 70-15-15 percent produced coefficients of determination equal to 0.99, and the Levenberg-Marquardt backpropagation function provided the best network performance. To help validate the model, the ANN was also used to evaluate chemical-disease relationships for three uncured chemicals. Results showed that ANNs have the potential to predict disease associations for uncured chemicals and to guide research for cured chemicals that may require further toxicological testing.</p>					
15. SUBJECT TERMS Artificial Neural Network (ANN), Predictive Model, Chemical-Disease Relationships, Chemical Structure, Chemical Classification					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	18. NUMBER OF PAGES 166	19a. NAME OF RESPONSIBLE PERSON Dr. Willie F. Harper, AFIT/ENV
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U			19b. TELEPHONE NUMBER (Include area code) (937) 255-3636, x 4528 Willie.Harper@afit.edu